Spring 2015

The effects of a planned missingness design on examinee motivation and psychometric quality

Matthew S. Swain
James Madison University

The Effects of a Planned Missingness Design on Examinee Motivation and Psychometric Quality

Matthew S. Swain

A dissertation submitted to the Graduate Faculty of

JAMES MADISON UNIVERSITY

In

Partial Fulfillment of the Requirements

for the degree of

Doctor of Philosophy

Department of Graduate Psychology

May 2015

Acknowledgements

Several individuals have been instrumental in the completion of this dissertation and my doctoral program. I would first like to thank my PhD advisor, mentor, and friend, Dr. Donna Sundre for her unwavering support over the past few years. Donna has been more than I could have hoped for in an advisor. Out of all of my professors in graduate school, she's the one I will miss the most. I would also like to thank the other members of my dissertation committee: Dr. Dena Pastor and Dr. Debbi Bandalos. Thank you, Dena, for putting up with my incessant questions and worried antics throughout my dissertation saga. I, quite literally, could not have done this without you. Thank you, Debbi, for your constant support through these past few years. You have shaped my professional career in many positive ways. I have enjoyed your inquisitive mind, kindness, and sense of humor.

I would also like to thank all of the faculty of the Assessment and Measurement program for their contributions, big and small, to my graduate school experience. I am proud to have been your student. Several fellow students, too many to list, have been my stronghold through the ups and downs of graduate school. Specifically, I want to thank (in no particular order) Bo, Laura, Mandy, Jason, Dan, and Liz for their support in school and in life. Finally, I am thoroughly indebted to the love and support I have received from my family. Thank you to my parents for being proud of me no matter how hard it is to describe my degree. I'm finally getting off the school bus, Dad. And Mom, even though I've learned a lot in school, you have taught me all of the important things about life and love. And to my new family: Meredith, you have been the best partner throughout this journey. You deserve an honorary degree for all of the support you have given me over the past few years. This is for you. I cannot wait to see where our life together leads.

Table of Contents

List of Tables

List of Figures

Abstract

Assessment practitioners in higher education face increasing demands to collect assessment and accountability data to make important inferences about student learning and institutional quality. The validity of these high-stakes decisions is jeopardized, particularly in low-stakes testing contexts, when examinees do not expend sufficient motivation to perform well on the test. This study introduced planned missingness as a potential solution. In planned missingness designs, data on all items are collected but each examinee only completes a subset of items, thus increasing data collection efficiency, reducing examinee burden, and potentially increasing data quality. The current scientific reasoning test served as the Long Form test design. Six Short Forms were created to serve as the planned missingness design that incorporated 50% missing data. Examinees mid-way through their educational career were randomly assigned to complete the test as either a Long Form or Short Form. Multiple imputation was used to estimate parameters for both conditions. Group mean test performance was higher in the Short Form condition compared to the Long Form. Although slightly higher examinee motivation and less shared variance between test-taking effort and test performance was observed in the Short Form condition, these effects were not statistically significant. Internal consistency coefficients and item parameter estimates were also similar between the form conditions. Although effect sizes were small, the implications of these results for assessment practice are substantive. This study supported the use of planned missingness designs for accurate estimation of group student learning outcomes without jeopardizing psychometric quality. The synthesis of planned missingness design and

examinee motivation literatures provide several opportunities for new research to improve future assessment practice.

CHAPTER ONE

Introduction

Assessment and accountability needs in U.S. higher education have increased in recent years because of heightened demands from policy-makers and the public. In 2006, the U.S. Department of Education brought national attention to insufficient higher education accountability efforts. The authors of this report painted a dour picture of the state of higher education, citing falling or stagnant literacy rates of present-day college graduates compared to citizens who graduated just a decade prior. Several recommendations were made by the Department of Education, one of which called for student learning assessment data to be collected and provided to the public as a demonstration of what a student may achieve by attending their institution. Since that time, higher education assessment endeavors have expanded to answer the call for data that demonstrate student learning as well as program and curriculum improvement (Ewell, 2009; Suskie, 2010).

In response to this increased demand for more data demonstrating learning gains, a test-based accountability system has become the trend in higher education across many states (Zis, Boeke, & Ewell, 2010). Several tests are available for assessment practitioners to implement such as the Collegiate Learning Assessment (CLA) or the National Survey of Student Engagement (NSSE). Other practitioners may design their own instruments to provide targeted feedback to the educational curriculum at their institution (e.g., scientific reasoning skills). Scores from these tests are used for important purposes: to demonstrate learning gains over time (i.e., accountability), and to inform changes to educational curriculum (i.e., assessment). These score-based inferences may

result in rewards or consequences that directly affect internal stakeholders, such as academic programs and student affairs offices, as they report to external audiences.

Scores from these assessment and accountability tests have implications for various parties who have a stake in higher education. Internal stakeholders, such as university administrators, department heads, program managers, and assessment practitioners, rely on assessment and accountability test scores to make curricular and program improvements as well as to justify institutional resource request and use. External stakeholders, such as regional and professional program accrediting agencies, will make judgments concerning institutional quality in part as a result of these assessment results. Furthermore, prospective students and parents potentially use test scores to make decisions about which institution will provide the greatest return on their educational investment. This concern stems from the rise in college costs. In the past ten years, the college tuition rate for in-state students attending public institutions has risen 42% and has more than doubled in the past 20 years (College Board, 2014). For university staff and prospective students, assessment and accountability tests are *high-stakes,* because demonstrated gains, or a lack thereof, directly and personally affect them.

Other stakeholders in higher education are not at risk for direct consequences from assessment and accountability tests. Chief among these stakeholders are the college students whose learning progress is being measured. When data are being collected for assessment and accountability purposes, it is not uncommon for there to be no personal consequences tied to an individual student's performance. Therefore, the same tests that are considered high-stakes to university administrators and staff are perceived as *low-stakes* by the students themselves (Cole & Osterlind, 2008; Smith & Smith, 2002). From

the student perspective, low-stakes testing contexts are easily distinguished from high-stakes testing contexts. For example, students clearly know that their personal SAT or GRE scores will be used for college and graduate school admissions decisions. These are high-stakes, because their *individual* scores result in a direct and meaningful consequence. Conversely, student performance on assessment and accountability tests do not result in any direct or personal consequences. Student performances are typically aggregated with those of their peers and reported at the institution level. Thus, individual student performances function in an anonymous and quite impersonal context.

We can clearly understand how empirical studies of low-stakes testing contexts have consistently shown that students perform better on tests with consequences than tests with no consequences (Sundre, 1999; Sundre & Kitsantas, 2004; Wolf & Smith, 1995). If students are underperforming on assessment and accountability tests due to the low-stakes nature of the exam, the validity of the inferences made from these scores is undoubtedly jeopardized. Learning gains may actually be occurring during college, but without consequences for performance, students may not be motivated to perform to their optimal ability. Test-taking motivation during a low-stakes exam becomes a key issue when considering the potential judgments and decisions that the internal and external stakeholders described previously will contemplate. Therefore, examinee motivation merits a great deal more attention during institutional testing.

Examinee Motivation

General motivation has been defined as the "energization and direction of human behavior" (Deci & Ryan, 1985, p. 3). Similarly, motivation in a testing context has been defined as a "student's engagement and expenditure of energy toward the goal of

attaining the highest possible score on a test" (S. L. Wise & DeMars, 2005, p. 2). Imagine watching a room full of students completing a test. The motivated examinees are attending to the test materials, carefully completing their answer sheets, and checking their responses for accuracy. The non-motivated examinees may answer a few questions but generally appear apathetic, only giving enough effort to complete their requirement. One might suspect that motivated examinees are more likely to perform better on a test when compared to unmotivated peers. Unfortunately, lower performance from unmotivated examinees has consistently been observed in low-stakes testing contexts. A positive correlation exists between examinee motivation and test performance in that examinees who try harder score higher on a test (Abdelfattah, 2010; Brown & Gaxiola, 2010; Huffman, Adamopoulos, Murdock, Cole, & McDermid, 2011; Silm, Must, & Täht, 2013; Wolf, Smith, & Birnbaum, 1995). In a review of the literature, S. L. Wise and DeMars (2005) noted that motivated examinees in educational assessment contexts performed better than unmotivated examinees by an average of 0.59 standard deviations. We might conclude that it makes sense that students would have higher motivation if they had greater ability to perform well. Importantly, the authors report that test-taking motivation is typically *not* correlated with student ability as measured by the SAT. That is, ability is not a common factor between motivation and test performance; rather students of all abilities who fail to give their best effort on an educational assessment test will perform below their true ability. This dilemma has many ramifications.

When test scores are lower than they should be due to a motivation issue, the interpretations based on those test scores will be affected (Wainer, 1993). Eklöf (2010) asserts that "not acknowledging student motivation in the assessment situation and the

impact of motivation on performance may pose a threat to the validity of the interpretation and use of assessment results" (p. 345). Validity, or the interpretation of test score meaning, is clearly affected when students are not motivated to perform well. Moreover, a complete lack of motivation may result in students skipping the testing session. Practitioners have noted abysmally low attendance rates for institutional assessment sessions, sometimes as low as 14% (Porter & Umbach, 2006). When data are collected from students who are not representative of the institution, the validity of score inferences is most certainly hampered.

Assessment practitioners who fail to measure examinee motivation during accountability testing may make unwarranted inferences about the meaning of test scores. Moreover, Standard 13.9 of *The Standards for Educational and Psychological Testing* recommend consideration of supplemental information to better interpret test scores. They specifically caution that the meaningfulness of tests to examinees will influence their motivation and impact test performance during low-stakes accountability testing contexts (AERA, APA, & NCME, 2014). Clearly, measuring and understanding the influence of examinee motivation on test scores should be an aim of education assessment practitioners who strive to report and appropriately interpret test scores. Test-taking motivation has been conceptualized and described in a few viable ways: the expectancy-value theory and the demands-capacity model.

Expectancy-value theory. Applying motivational theory to the examinee experience may help clarify how an examinee determines his or her motivation and suggest solutions to alleviate its negative influence. Expectancy-value (EV) theory is one framework that demonstrates how an individual makes decisions concerning the level of

effort or motivation he or she is willing to invest in achievement situations (Eccles & Wigfield, 2002; Wigfield & Eccles, 2000). According to the theory, an individual generally considers two factors to make this decision: how well they *expect* to perform on the task and the *value* the task provides.

Level of expectancy in achievement situations has been defined succinctly as "probability of success" (Eccles (Parsons) et al., 1983, p. 81). Expectancies are determined in part by an individual's performance beliefs, or how well they imagine they will perform. Individuals typically have some experience with the content of an impending task and can draw from this experience to determine their anticipated success. For example, students in a math class who are about to take a final exam consider how well they have performed on previous class assignments and tests to determine their expectations regarding their final exam performance.

Task value may be conceptualized as whether the individual perceives the task to be worthwhile. Task value encompasses (a) attainment value, (b) intrinsic value, (c) utility value, and (d) cost (Eccles (Parsons) et al., 1983; Wigfield & Eccles, 2000). Students in the previous math class example may consider high performance as valuable for different reasons. Individuals who want a high score on their final exam would be considered to have high attainment value if they want a good final grade. If math is inherently interesting to the student, the test has high intrinsic value. Utility value in this example refers to how getting a good score will be beneficial to one's future. Thus, for many students completion of this prerequisite math course will unlock entry to more inherently interesting and personally important courses. Finally, motivational or psychological cost is "what the individual has to give up to do a task…as well as the

anticipated effort one will need to put into task completion" (Wigfield & Cambria, 2010, p. 4). Some theorists view motivational cost as a forgotten component of EV theory deserving of more attention in research (Barron & Hulleman, 2015). Essentially, cost encompasses the negative aspects of completing a task such as denying alternatives (i.e., spending time with friends instead of homework) or how much effort the task might require. For example, an arduous or lengthy essay final exam would be considered to have higher cost than an easier to complete, shorter multiple-choice exam.

Through the lens of EV theory, an examinee's motivation during a test is determined by their expectancies about their test performance and how much they value doing well on the test. These constructs are theoretically important as task value may be conceptualized as a predictor of test-taking effort, which is then a predictor of test performance (Cole, Bergin, & Whittaker, 2008). In a high-stakes testing context such as the SAT, students place high value on a high score, and therefore, expend the effort necessary to achieve the highest possible score. In low-stakes testing contexts, doing well on a test often has little perceived importance or value for examinees, and therefore, no attainment, intrinsic, or utility value to most examinees (Wise & DeMars, 2005). Furthermore, the costs of task completion, particularly for challenging assessment tests, may render the context as unworthy of such meaningful engagement. Low value may affect test-taking effort and subsequently result in poor test performance in low-stakes testing contexts. Indeed, examinee test-taking effort has also been shown to relate more strongly with examinee performance for low-stakes tests when compared to high-stakes tests of similar difficulty (Smith & Smith, 2004). This means that student performance on two exams of comparable difficulty was better predicted by their individual effort in a

low-stakes context than a high-stakes context. The relationship between test value, test-taking effort, and test performance is one that has been frequently discussed in the examinee motivation literature. These relationships will play a central role in this dissertation study.

EV theory applied to the testing context can also help explain how an examinee *approaches* an exam and how examinee motivation may change as a result of their *experience* when completing the exam. For example, an examinee may expect to do well on a test of their favorite subject and believe that doing well will positively benefit their university. Thus, they begin the test with high examinee motivation. However, once this examinee encounters difficult or complex items, their motivation may change due to a change in perceived expectation and/or cost. It is conceivable that examinee motivation is a dynamic process that is borne out of an examinee's interaction with a test and the testing context. Similar to EV theory, the demands-capacity (DC) model provides an explanation for examinee motivation (S. L. Wise & Smith, 2011).

Demands-capacity model. The DC model presents examinee motivation as a dynamic process where an examinee may change their motivation depending upon a number of factors in the process of completing a test. Examinee motivation is determined by two constructs: (a) *resource demands* and (b) *effort capacity*. Resource demands relate to the effort required to complete an item. This construct is considered fixed for any item; however, different items may have different levels of resource demands, and these demands will accumulate over the test taking experience. Effort capacity is the level of effort an examinee is willing to expend on a test. Effort capacity may vary across examinees but may also be renegotiated within an examinee as they experience test

engagement. According to this model, resource demands and effort capacity are malleable within individuals as they experience and react to testing conditions.

Since effort capacity will vary within and between examinees, the factors that determine this model construct may be manipulated. The goal is to increase effort capacity so an examinee will perform to the best of their ability throughout the testing experience. Factors that determine effort capacity include external factors such as the test consequences (i.e., the stakes of the test) and internal factors such as fatigue (i.e., many difficult prior items) and test anxiety among others (see Wise & Smith, 2011). The DC model lends additional explanatory power to the discussion of examinee motivation during a low-stakes exam.

Several similarities exist between the EV and DC models but one in particular is the concept of fatigue or task cost. The costs associated with task completion are particularly relevant in low-stakes testing contexts. Motivational cost involves any negative component of the task (Eccles & Wigfield, 2002). In a testing situation, examinees may find the test arduous (i.e., mentally taxing) in content or length. The exam may also be viewed as a poor alternative to spending time with friends or doing homework. An increase in motivation cost results in a decrease in perceived value of the task. With lower value, test-taking effort may diminish resulting in lower test performance. Assessment and accountability tests measure constructs that are deemed important to the university but have varied importance to students. Moreover, some of these assessments are lengthy and mentally taxing, both of which increase the costs associated with taking a test that may be personally irrelevant. The EV and DC models

provide important frameworks that illuminate examinee dynamics known to influence test performances. All of these jeopardize intended test interpretations.

Assessment practitioners must balance the issues of motivation in low-stakes testing contexts with the increasing demands for evidence of student learning growth and development. Institutional data needs must be met. These competing demands must be considered carefully. A reduction in the number of tests is not considered a reasonable solution; in fact, for many accountability programs, the content and selected tests are mandated. Fortunately, several alternative data collection designs exist for consideration. Data collection designs specify how data are to be collected for assessment or research purposes. One might assume that to collect desired information all participants must complete all items or tests. Consideration of the EV and DC models of test-taking motivation indicates that such a design would not produce high quality data. In an institutional assessment context, not all items have to be administered to all participants. This is because individual scores are not the intended measurement level. In most cases, students are nested within the desired measurement level: classrooms, programs, institutions, districts, states, etc. Measurement at the program or institution level can be accomplished through *planned missingness designs*.

Planned Missingness Designs

Planned missingness designs (PMDs) are a family of data collection designs that assist practitioners and researchers in collecting all relevant information to make group-level inferences (Graham, Taylor, Olchowski, & Cumsille, 2006). These designs may also be referred to as *efficiency designs* in the literature. The terminology may differ depending on the context within the field but the general concept is the same: all items

are divided into sets or blocks that are systematically combined into new forms. The intent is to create different forms comprised of unique item set combinations. Some forms contain some item sets while excluding others. Through these means, no student would be asked to complete all item sets, thus the time and effort required would be diminished. These forms are then randomly assigned and administered to groups of participants. Whether to increase the efficiency with which information is collected (i.e., efficiency designs) or to plan missingness for some items to make room for others (i.e., PMDs), these designs present a potential solution to reduced examinee motivation. For simplicity in the presentation of the varied designs available to practitioners, all designs reviewed will be referred to as PMDs.

PMDs do not require all students to complete all measures and/or items to make institutional-level decisions about student learning or program effectiveness. PMDs involve the strategic incorporation of missing data to shorten the testing time or to provide the time to collect more information. Thus, PMDs may provide important clues for a solution to low examinee motivation via a reduction in examinee cost (i.e., shorter tests) while simultaneously allowing assessment practitioners to collect data on every mandated and desired construct. It should be noted that PMDs are often conducted on the test-level where data are collected on multiple tests, yet each student only completes a subset of randomly assigned instruments rather than items. However, item-level sampling procedures will be the focus of discussion for this dissertation. There are several item-level designs available to assessment practitioners and researchers. Common PMDs include but are not limited to: (a) matrix sampling; (b) balanced incomplete block; and

the (c) three-form design. Each will be described in turn followed by a comparison of the three designs.

Matrix sampling design. Matrix sampling involves the creation of multiple forms of a test from subsets of the total bank of items (Shoemaker, 1973). First, a large collection of items is divided into smaller item subsets. These item subsets are then combined into test forms or given as forms themselves to a random sample of students from the population of interest. This practice of sampling both items and students is referred to as *genuine matrix sampling* (Popham, 1993). Matrix sampling designs are fairly common in large-scale testing programs at the state-level in the U.S. with examples such as Kentucky, Maine, Maryland, Massachusetts, Oregon, Pennsylvania, and Wyoming, and internationally in the Dutch National Assessment Program (DNAP; Childs & Jaciw, 2003). These large-scale testing programs collect information about a multitude of constructs, and matrix sampling provides an efficient way to collect all desired information without unreasonable examinee burden. As an illustrative example, consider a testing program that collects data using a 100-item test. The practitioner could divide this 100-item test into four, 25-item forms and then randomly assign a fourth of the total sample to complete one of these shortened test forms. In this way, data are collected on all test items but individual examinees only complete 25 items rather than 100 items. Figure 1a contains an example of matrix sampling within a single test.

Balanced incomplete block design. The balanced incomplete block (BIB) design is a more complicated form of matrix sampling (Johnson, 1992). This sampling design has seen wide use in the U.S. with the National Assessment of Education Progress (NAEP; Zwick, 1991). Like matrix sampling, the BIB design divides items into *blocks*

Test Form	Item Set A (k = 25)	Item Set B (k = 25)	Item Set C (k = 25)	Item Set D (k = 25)
Form A	I	M	M	M
Form B	M	I	M	M
Form C	M	M	I	M
Form D	M	M	M	I

Figure 1a. An illustrative example of a matrix sampling design using a 100-item test. "I" indicates item set is included on the form. "M" indicates item set is missing. *k* is the number of items in each set. In this example, one-fourth of the total sample of examinees could be randomly assigned to complete one of the four forms.

	Blocks						
Booklet	A	B	C	D	E	F	G
1 – ABD	1	2	M	3	M	M	M
2 – BCE	M	1	2	M	3	M	M
3 – CDF	M	M	1	2	M	3	M
4 – DEG	M	M	M	1	2	M	3
5 – EFA	3	M	M	M	1	2	M
6 – FGB	M	3	M	M	M	1	2
7 – GAC	2	M	3	M	M	M	1

Figure 1b. An illustrative example of a balanced incomplete block (BIB) design using seven blocks. BIB designs follow three rules: (a) each block is completed in the same amount of time; (b) each block is presented in each order position; and (c) each block is presented with all other blocks on one form.

Test Form	Item Set X (k = 25)	Item Set A (k = 25)	Item Set B (k = 25)	Item Set C (k = 25)
Form XAB	I	I	I	M
Form XBC	I	M	I	I
Form XAC	I	I	M	I

Figure 1c. An illustrative example of the three-form design using a 100-item test. "I" indicates item set is included on the form. "M" indicates item set is missing. *k* is the number of items in each set. In this example, one-third of the total sample of examinees could be randomly assigned to complete one of the three forms.

that may vary in item number but will require the same time to complete. BIB designs follow three rules: (a) each block is constructed to be completed in the same amount of time; (b) each block is presented in all possible order positions; and (c) each pair of blocks is presented on at least one form. Two or three of these blocks are assembled into booklets. The requirement that each block appears in each position and is paired with all other blocks results in a "balanced" design. This design is also "incomplete" due to the lack of some blocks in each booklet. Figure 1b depicts an illustrative example of the BIB design with seven blocks.

Three-form design. Another common design used to increase efficiency in data collection efficiency is the three-form design (Graham, Taylor, & Cumsille, 2001). In this design, items are split across three or more item sets and then combined into various forms. As seen in the illustrative example in Figure 1c, item sets are combined in a way that allows all possible pairs of items to be presented on at least one form much like the BIB design. However, the three-form design differs from the BIB and matrix sampling by the inclusion of an optional common set of items across all forms called the X set. The X set may contain items that are crucial to the research study or may be used to collect information that is simply required from all examinees. X set items may be placed first in the forms; however, this is not necessary. Researchers may create as many item sets and forms as they would like to incorporate more missingness. In the illustrative example in Figure 1c, which is a typical three-form design, each form contains 75 items, or 25% missingness when compared to the full item pool. This means that all examinees will complete 75% of the total number of items. Other researchers have incorporated much

more missingness in PMDs, up to 60% missingness in some cases (Raghunathan & Grizzle, 1995).

Another key difference between matrix sampling, the BIB design, and the three-form design is the ability to correlate item responses between blocks. For example, consider how items 1 and 26 are collected in a matrix sampling and a three-form design (see Figures 1a, c). If we assume that items 1 through 25 appear in the first item set and 26 through 50 appear on the second, items 1 and 26 are never administered to the same student in a matrix sampling design. In the three-form design, items 1 and 26 are both presented to students who complete Form XAB. Regardless of which block items 1 and 26 appear in the BIB design, by definition these two blocks will be presented together on one booklet. Because all pairs of items are presented in the BIB and three-form designs, all correlations between items can be estimated (Graham, Hofer, & MacKinnon, 1996). When researchers or assessment practitioners are interested in relationships between items from different item sets, the BIB and three-form designs are deemed superior to the matrix sampling design, because these item covariances can be calculated.

Planned Missingness vs. Complete Tests

There are both pros and cons to implementing a PMD instead of a full-form design where all items are administered to all participants (Graham et al., 2006). PMDs are typically presented as efficiency designs, because they allow for the collection of more information than may be afforded to the practitioner or researcher. For example, if an assessment practitioner is interested in collecting data on 60 items, but is only allowed 45 minutes to administer these items, a three-form design could be used to create multiple forms that contain 40 items each. Without an X set, items could be divided into three sets

of 20 items each: Form AB would contain items 1-40, Form BC would contain items 21-60, and Form AC would contain items 1-20 and 41-60. When resources such as examinee time are limited, PMDs provide an efficient method for assessment practitioners to collect all necessary data. However, this design does not come without costs.

In all PMDs, participants are randomly assigned to complete one test form; this reduces the sample size for any one item when compared to a full-form design. Random assignment results in about equal sample sizes completing each form or booklet; however, the exact sample size for each parameter will differ depending on which PMD is implemented. These sample sizes will always be smaller than in a full-form design where all participants complete all items. Table 1 describes the different expected sample sizes for item means and item correlations within each PMD discussed. With smaller sample sizes for any parameter, the standard error of that parameter increases and the statistical power of any desired significance tests decreases. These smaller sample sizes are estimated after applying *pairwise deletion* where only the cases that contain complete data on both items are retained for estimation. Pairwise deletion is one of the traditional missing data techniques available to practitioners; though used extensively, it is not typically recommended (Peugh & Enders, 2004). Fortunately, advances in missing data handling techniques have resulted in methods that can recover missing information (Graham, Hofer, & Piccinin, 1994). These methods have been shown to allay the trade-offs associated with PMDs, such as lower statistical power. The *missing data mechanism* or the reason behind data being missing (D. B. Rubin, 1976) is within the control of the researcher when implementing a PMD. Most missing data handling techniques assume a

Table 1
Comparison of Planned Missing Data Designs using a Fixed Sample Size (N = 1,000)

	Item Means Estimable?	Item Correlations Estimable?	Sample Size For Each Item Mean (N = 1,000)	Sample Size For Each Item Correlation (N = 1,000)	Percent of Total Items Completed by Any One Examinee
Long Form	Yes	Yes	1,000	1,000	100%
Matrix Sampling	Yes	No [a]	250	0 [a]	25%
Balanced Incomplete Block (BIB) Design	Yes	Yes	≈ 429	≈ 143	50%
Three-Form Design with X Set	Yes	Yes	X Set = 1,000 A, B, & C Set ≈ 667	XX [b] = 1,000 AA, BB, CC ≈ 667 XA, XB, XC ≈ 667 AB, BC, AC ≈ 333	75%
Three-Form Design without X Set	Yes	Yes	A, B, & C Set ≈ 667	AA, BB, CC, AB, BC, AC ≈ 667	≈ 67%

Note. All scenarios applied the designs depicted in Figure 1 to a hypothetical sample of 1,000 participants. All sample sizes are computed using pairwise deletion.
[a] Item correlations are not estimable with the matrix sampling design unless those items are within the same item set.
[b] Denotation for sample sizes in the three-form designs considers the placement of the variables in each item set. For example, a correlation between one item in item set A and one item in item set B would be an AB effect.

missingness mechanism, and violations of this assumption can bias parameter estimates and standard errors in known ways.

Missing Data Mechanisms

Educational assessment data may be missing for many reasons. As discussed above, data may also be missing by design. The missingness mechanism explains why data are missing from a data set. D. B. Rubin (1976) coined the terms most commonly used to define missing data mechanisms. In D. B. Rubin's view, missingness (R) is a binary variable (1 = observed, 0 = missing) that has a probability distribution. Although it is impossible to know the precise probability distribution behind R, other variables in the data set may relate to the probability of missingness, and it is these relationships that define the missing data mechanism (Enders, 2010).

When data are missing completely at random (MCAR), the missingness (R) is neither related to the variable of interest (Y) nor to any other variable in the analysis (X). In other words, missingness is not related to the would-be values of the missing variable nor other variables in the dataset. For example, in a study of substance abuse, surveys sent to some participants may get lost in the mail. Responses from study participants who never received their survey would be missing data completely at random, because missingness (R) is considered unrelated to any of the variables they would have completed (X or Y).

Data may also be missing at random (MAR), which differs from MCAR despite including the term "random." In a MAR scenario, variables in the dataset (X) may be related to the missingness mechanism (R), but missingness is not related to the missing variable (Y) after controlling for the relationship between X and R. For example, in the

same study of substance abuse, participants with extreme scores on a pretest survey (X) may be referred to a treatment program by surpassing a cut-score. If a posttest measure is collected (Y) from participants in the treatment program but not from all participants in the study, then the missingness mechanism (R) is not related to the posttest scores (Y) after controlling for the relationship between pretest (X) and the missingness (R). In other words, data on Y are missing only for participants who scored below the cut-score at pretest (X).

Data may be missing not at random (MNAR) which is also referred to as non-ignorable missingness (Allison, 2009). MNAR data arises when the relationship between the missingness mechanism (R) and the missing variable (Y) exists even after controlling for the relationships between other variables in the analysis (X) and the missingness (R). In the substance abuse study example, data may be MNAR if some participants failed to report their substance abuse because their level of use was so extreme as to be socially undesirable. Notably, if a variable was introduced to the analysis model that explained this MNAR data, the missingness mechanism may change to a MAR mechanism. Continuing with the substance abuse example, if a measure of social desirability was collected, scores on this variable may relate to missingness (R) if participants were failing to report their use for social reasons. That is, participants who score high on the social desirability instrument may be more likely to fail to report their usage level. If social desirability scores were included in the analysis, the missingness mechanism may satisfy the MAR assumption rather than the MNAR assumption. This example should also illuminate how missingness mechanisms are a property of an analysis model rather than a property of an entire dataset (Baraldi & Enders, 2010). MNAR data can be troublesome

for researchers to handle requiring advanced statistical models that make strict assumptions to result in unbiased parameters. However, PMDs produce a main missing data mechanism that is MCAR by design, which is much easier to overcome than MNAR. For this reason, the missingness mechanism (R) is known and it is not related to the item score (Y) or any other variable in the dataset (X).

We know that PMD data is MCAR due to the random assignment of participant to test form. Therefore, we do not need to test this assumption although a test does exist (R. J. A. Little, 1988). When data are collected via a PMD such as the three-form design, any analysis that involves variables that contain planned missing data are affected by the missingness mechanism imposed by the PMD, which is MCAR. However, this property affords assessment practitioners and researchers with certain expectations regarding the results of analyses that include missing data. With the use of advanced missing data handling techniques such as *multiple imputation* and *maximum likelihood* estimation, missing data can be "recovered" using statistical software.

Briefly, both multiple imputation (MI) and maximum likelihood (ML) estimation may be employed to recover missing data. These methods both assume the MAR missingness mechanism underlies missing data in the analysis and have been shown to result in unbiased parameters and standard errors when compared to traditional deletion techniques (Enders & Bandalos, 2001; Enders, 2001; Graham et al., 1996; T. D. Little, Jorgensen, Lang, & Moore, 2014; Schlomer, Bauman, & Card, 2010). Additionally, both MI and ML are now readily available in several commercially available and free software packages (Peng, Harwell, Liou, & Ehman, 2006). The development and availability of these advanced procedures should provide assessment practitioners with the assurance

that PMDs could improve the efficiencies of assessment data collection without a reduction in psychometric integrity. If the results of the current study find few deleterious effects on data quality when using a PMD, assessment practitioners could use PMDs to reduce testing time for each construct while sustaining data quality. Moreover, if shortened test forms improve test takers' experience by reducing examinee burden or cost, PMDs could provide an innovative approach to addressing examinee motivation in low-stakes testing contexts and therefore result in *improved* data quality.

Addressing Examinee Motivation with a PMD

Low motivation is problematic during educational assessment testing contexts given the known and previously described deleterious relationship between examinee motivation and test performance. The current study attempts to bridge two previously distinct research areas. More specifically, the study blends relevant examinee motivation literature with planned missingness literature to pose the question: does a PMD result in data of equal or better quality than a traditional, full-form design? PMDs reduce test length by creating multiple test forms and randomly assigning examinees to a single, shortened form. According to both EV theory and the DC model, a reduction in test length, which translates into a reduction in examinee burden or cost, may result in an increase in test-taking effort. That is, a PMD may increase test-taking effort enough to *decrease* the influence examinee motivation has on educational assessment test scores. Researchers in survey methodology note that "shorter questionnaires lead to better response rates and less survey fatigue near the end of the questionnaire, increasing data quality" (Littvay, 2009, p. 112). It remains to be seen if the same positive effect on data quality will occur during higher education assessment and accountability testing contexts.

One of several purposes of the present study is to explore whether examinees perform similarly on a single test when a PMD is compared to a complete form. Related to the discussion of planned missingness, modern missing data handling techniques used to "recover" data missing by design are outlined. These techniques are presented to demonstrate the ease of analyzing data collected via a PMD and the advantages of these modern techniques over traditional approaches. With a bit of advanced planning to design a PMD and the use of an advanced missing data handling technique, assessment practitioners may be encouraged to use more PMDs when collecting assessment data. These efficiency designs will be of much greater interest and utility if they are also shown to improve the measurement precision and validity of student learning in higher education.

CHAPTER TWO

Literature Review

Chapter 1 defined planned missingness and examinee motivation and in this chapter, I use those terms to review the literature relevant to the current study. First, I review several prior interventions to curb the effect of low examinee motivation during low-stakes testing contexts. The benefits and costs of each intervention are presented to demonstrate the need for an intervention that addresses examinee burden, or motivational cost, and its influence on group test scores. Next, item position effects are discussed as high examinee burden (i.e., an arduous test) may have deleterious effects at the item level (i.e., increased difficulty). Administering the same test using both a full-form design and a PMD provides a unique opportunity to examine differences in item-level characteristics. Finally, missing data handling techniques are reviewed. Although PMDs serve to reduce examinee burden, the planning to exclude data can affect sample sizes. When using a traditional deletion technique, sample sizes are reduced along with statistical power. Modern missing data handling techniques are available to overcome these issues and "recover" data collected via a PMD. This literature review is organized to discuss: (a) prior interventions to improve examinee motivation, (b) possible item-level differences between a PMD and a full-form test, and (c) missing data handling techniques that result in unbiased parameter estimates and standard errors.

Methods to Curb Low Examinee Motivation

The studies reviewed here cut across secondary and higher education yet all pertain to low-stakes testing contexts. Motivational theories, like EV theory (Wigfield & Eccles, 2000) and the DC model (S. L. Wise & Smith, 2011), suggest several factors that

should directly translate into increased motivation during a task. For example, drawing examinee attention to the value of their performance (i.e., scores are used to improve their university) may increase effort and subsequent test performance. Another possible factor would be to tie consequences to test performance such as awarding points toward a course grade. Researchers have attempted to manipulate these and several other factors in the testing environment to improve low examinee motivation in low-stakes contexts. There are several techniques that have been proposed to increase examinee motivation. The studies reviewed here focus on intuitive possibilities to increase test value and consequences by (a) informing students of assessment test score use, (b) tying performance to grades, (c) promising monetary incentives, and (d) providing test score feedback.

Test score use. If low examinee motivation is frequently observed during low-stakes exams, one apparent solution would be to change the consequences of the test to raise the stakes. By raising the stakes of performance, examinees increase in their perceptions of test importance and in turn expend sufficient effort to perform to the best of their ability (Cole et al., 2008). Assessment practitioners use test scores for programmatic and curricular decisions, as well as accountability mandates. Perhaps informing students of these important uses would impact examinee motivation. Liu, Bridgeman, and Adler (2012) proposed that increasing the consequential use of test scores should increase examinee effort and subsequent performance. The researchers designed three conditions: (a) the Control condition informed students that their scores would not be disclosed and would be used for research purposes, (b) the Institutional condition told students that their scores were to be used at the aggregate level to evaluate

the quality of instruction at their institution, and (c) the Personal condition in which students were told that their personal scores may be released to faculty and potential employers to evaluate their individual ability. College students in the Control condition scored significantly and practically lower than the Institutional and Personal conditions on the ETS Proficiency Profile; however, there were no differences in performance between the Institutional and Personal conditions. A similar effect was observed for total test-taking motivation scores on the Student Opinion Scale (SOS; Sundre & Moore, 2002). Students in the Personal and Institutional conditions reported higher mean test-taking motivation than those the Control condition, but these conditions were not different from each other. Although significant differences in test performance and test-taking motivation were observed by Liu and colleagues, observed differences may be a function of test instruction misalignment with typical higher educational assessment contexts. More specifically, practitioners collect assessment data to make inferences and to report about student learning to a variety of stakeholders, not for research purposes as the Control condition stated. Therefore, the only realistic conditions relevant to educational assessment were the Institutional and Personal conditions between which there were no effects.

To test whether realistic testing consequence conditions had an effect on value-added estimates, Finney, Sundre, Swain, and Williams (2015) tracked a cohort of students from when they entered a university to mid-way through their sophomore year. In their study, first-year students were randomly assigned to one of three conditions where instructions for completion of a test of scientific reasoning were systematically changed. In the Institutional condition, consistent with the Liu et al. (2012) study,

students were told that their scores were to be averaged and reported at the institutional level. In the Feedback condition, test instructions were the same as in the Institutional condition and added a sentence that their scores and interpretive information would be available to them later in the semester. And finally, in the Personal condition, test instructions built upon the Feedback condition and also told students that their personal scores were going to be available to faculty. These same first-year students were then assigned to the same test consequence condition after completing between 45-70 credit hours.

Unlike Liu et al. (2012), the authors found that the test instructions did not affect change in test scores or change in effort over time (Finney, Sundre, Swain, & Williams, 2015). However, the authors noted that examinee *change* in test-taking effort was related to value-added estimates in scientific reasoning, with value-added estimates being the change in test performance for the same examinee over time. That is, a change in test-taking effort over time correlated with a change in test performance over time. Students in all three conditions increased in mean performance on the test with standardized effect sizes ranging from .52 to .68. However, mean test-taking effort decreased over time with standardized effect sizes ranging from d = -.63 to -.70. In a bivariate sense, change in test-taking effort correlated moderately with value-added estimates in all three conditions (r = .48 to .65) indicating that smaller changes in effort were related with higher gains in scores (and correspondingly, larger changes in effort were related with smaller gains in scores). Moreover, these relationships were not moderated by test instruction condition. In summary, attempts to enhance perceived test value resulted in nonsignificant differences across realistic assessment conditions. While significant growth in

performance was observed over time, the negative change in effort biased performance estimates. These are critical findings that speak to the importance of enhancing examinee motivation.

Another intervention strategy that has been suggested to increase examinee motivation involves a student's sense of academic citizenship. By making students aware that their performance contributes to the overall performance of their school, and that overall scores are used to improve their institution, their academic citizenship is engaged. Readers will recall from Chapter 1, that a sense of academic citizenship is one of the internal factors hypothesized to predict effort capacity in the DC model (S. L. Wise & Smith, 2011). Two experimental groups were designed to manipulate test instructions (Kornhauser, Minahan, Siedlecki, & Steedle, 2014). In one group, students were instructed that they were representing their university by taking the exam whereas the other group was told that they would have no personal consequences for their performance. Both groups completed the Collegiate Learning Assessment (CLA; Klein, Benjamin, Shavelson, & Bolus, 2007) and the Student Opinion Scale (SOS; Sundre, 1999). Unfortunately, no differences in mean test performance, test-taking effort, or test importance were observed. These results disagree with Wise and Smith's (2011) suggestion that a sense of academic citizenship should increase effort capacity. However, Wise and Smith also suggest increasing test consequences such as tying test performance to grades should increase examinee motivation.

Tying test performance to grades. When students are informed that their performance will influence their course grades, they tend to perform better and report higher test-taking motivation than students who are not given these instructions. In one

study, undergraduate students were randomly assigned to one of four testing conditions (Wolf & Smith, 1995). All students completed two equivalent forms of a test that were assembled using item analysis data so that the forms were of similar content and difficulty. One test contained the words "This test counts for your grade" at the top, whereas the other test stated "This test does *not* count toward your grade." Trained proctors also drew student attention to this difference between the forms. The order and consequential assignment for the forms was counter-balanced, resulting in four conditions. All participants completed a motivation questionnaire after completing their assigned tests. Participants reported significantly higher mean test-taking motivation for the consequential test when compared to the non-consequential test; the researchers observed a standardized effect size of 1.45 standard deviations. A main effect for test consequence was significant in that participants performed better on the consequential test compared to the non-consequential test. There was no main effect for experimental group (i.e., form order) or an interaction between group and consequence. This main effect for test performance resulted in an effect size of 0.26 standard deviations. Test-taking motivation scores were positively correlated with test performance in both the consequential (r = .351) and non-consequential (r = .232) conditions. It is noteworthy that these students were prepared to take this scheduled course examination. Importantly, the relationship between motivation and performance was stronger in an absolute sense when the test counted toward a grade than when performance bore no consequence.

Sundre (1999) replicated the consequential conditions of the Wolf and Smith (1995) study with a separate sample of undergraduate psychology majors. Two parallel forms were created with 30 multiple-choice items and one essay each. These two forms

were counterbalanced for order and consequential nature resulting in four conditions much like the former study. Participants completed a revised version of the motivation questionnaire containing 10 items that measure two separate but related dimensions of test-taking motivation (Sundre, 1997). The Importance subscale measured the examinee's perception of test importance to them (e.g., "Doing well on this test was important to me") whereas the Effort subscale measured expended energy and persistence during the test (e.g., "I gave my best effort on this test"). Both subscales consisted of five items. This revised motivation scale was completed four times: after the multiple-choice and essay sections of each test consequence condition.

Total motivation scores were created by summing Effort and Importance subscale scores which in turn were summed responses to the five Effort and five Importance items, respectively. Similar to the Wolf and Smith (1995) study, the consequential condition reported significantly *higher* test-taking motivation on both the multiple-choice portion (d = 0.79) as well as the essay portion (d = 1.59). Analyzed separately, the Importance subscale means differed largely between the conditions with an effect size of 1.08 whereas the Effort subscale mean difference was .59, a more moderate effect. The consequential condition resulted in higher reported Importance and Effort scores. However, unlike the former study, the correlation between total test-taking motivation and test performance was larger in an absolute sense in the non-consequential condition (r = .38) than in the consequential condition (r = .15). Participants again performed better in the consequence condition than the non-consequence condition on both the multiple-choice portion (d = 0.62) as well as the essay (d = 1.38). Notably, as these effect sizes make very clear, differences in performance on the essay portion of the test was much

greater between the conditions than the multiple-choice portion suggesting that test-taking motivation may be moderated by task difficulty. When tests have no consequences, such as those during low-stakes testing contexts in higher education, the deleterious effect of low motivation on test performance appears to be more pronounced during tasks of high difficulty. Adding test consequences, specifically by tying test performance to course grades, appears to allay this effect.

Although tying test performance to grades appears to result in high motivation and performance for U.S. students, an analogous effect has not been observed for German students (Baumert & Demmrich, 2001). Several conditions were examined simultaneously in their study. In one condition, the researchers told a sample of ninth grade secondary school students that their math grade would be affected by their performance on a sample of math items from the Programme for International Student Assessment (PISA). Participants in the "control" condition were given routine PISA instructions that told students their scores would contribute to an effort to assess math ability around the world. Participants completed several other relevant scales: one measured task value and was completed before the test. Another scale administered before and after completing the PISA items measured intended and expended effort. Students reported similar mean task value and intended/expended effort across conditions. Concerning performance on the PISA items, no treatment effect was observed. Unlike the previous two studies (Sundre, 1999; Wolf & Smith, 1995), students performed similarly at the mean level regardless of the test instructions tying test scores to grades. A lack of a treatment effect in the PISA study could be a result of all students performing well to either (a) contribute to their grades or, for the control condition, to (b)

assure that Germany's scores fare well in international comparisons. That is, students in the "control" condition may have been motivated to contribute to the study.

With the exception of the previous study, tying low-stakes test performance to grades appears to increase motivation and test performance. However, there are some downfalls to this practice that should be considered. Although informing students that their performance will influence their grades is an inexpensive treatment to implement, assessment practitioners may lack the prerogative to tie test performance to student grades. Further, some institutional student learning objectives may not be tied to a particular course and not all college students enroll in the same courses. For example, critical thinking skills, while beneficial to assess, are not easily tied to a particular course that all students complete. If grades could be influenced, the grade selected would surely influence motivation differentially. The amount of extra credit promised may also be a factor for students. Instead of tying test performance to grades, other examinee motivation interventions that might more broadly impact higher education students have been sought and studied.

Monetary incentives. Instead of receiving a grade, providing performance-based monetary incentives during low-stakes testing has been explored with mixed results. O'Neil, Sugrue, Abedi, Baker, and Golan (1992) created three conditions varying test instructions given before eighth and twelfth grade students completed the NAEP test. The conditions were as follows: (a) students were promised $.50 for every correct answer, (b) students were promised $1.00 for every correct answer beyond chance level, and (c) students as a class were promised $16.00 if the whole class exceeded a standard. No differences in average test performance were found among these conditions. In a

subsequent study, O'Neil, Sugrue, and Baker (1995) varied test consequences prior to a sample of NAEP math questions including a condition where students were promised $1.00 for every correct answer. A sample of eighth graders and a sample of twelfth graders from the U.S. were assigned to each experimental condition separately. The eighth grade students in the financial gain condition reported higher effort and performed better than in the other experimental conditions and the control group who were given standard NAEP instructions. However, a similar effect was not found for twelfth grade students in the same condition compared to their peers.

In a follow-up study, twelfth grade students were again promised performance-based financial incentives, but with more money (O'Neil, Abedi, Miyoshi, & Mastergeorge, 2005). Students were promised $10.00 for every correct answer on items from the Third International Mathematics and Science Study (TIMSS) because the authors surmised that the amount of money offered in previous studies was paltry for seniors in high school. However, there was no main effect for this monetary incentive as students promised performance-based payment performed similarly to students who did not receive these instructions. Effort was measured by six self-report items (e.g., "I worked hard on the math test"). The incentive group did report higher effort on the test compared to the control group; however, effort was not highly correlated (r =.10) with math performance across the sample resulting in similar performance between the groups.

Although monetary incentives have resulted in inconsistent effects for motivating U.S. students, perhaps the same intervention would function differently in other countries. As another condition the PISA study in Germany, Baumert and Demmrich (2001) offered one group of ninth grade students ten Deutsche Marks if they scored

higher than expected. The group of students who received standard operational PISA test instructions served as the control. Participants in the control and monetary incentive groups reported the same level of intended/expended effort and performed similarly on the PISA math items. Although monetary incentives do not appear to affect examinee motivation or performance consistently across grades, the use of this practice also has concerns.

For instance, paying students for test performance may be considered unseemly by the public, and universities might be criticized for this practice (S. L. Wise & DeMars, 2005). Even if paying for performance was found to be effective and acceptable, the funds would need to be budgeted somewhere in the university. This may be difficult for some institutions, and for most institutions this simply would not be a sustainable practice. Moreover, all studies reviewed here concern K-12 assessment and a comparable intervention has not been conducted using college student samples. As alluded to previously, other researchers have attempted more appropriate incentives to enhance motivation and performance. It has been suggested that promising students interpretive test performance feedback would bolster motivation and performance while avoiding the practice of paying students to perform well on a test (S. L. Wise, 2009).

Feedback. Providing interpretive test score feedback about educational assessment performance has been explored. In addition to a condition where grades were tied to performance and a monetary incentive was provided, the study by Baumert and Demmrich (2001) also included a condition where students were promised feedback from their math teacher. Like the other experimental conditions, the students in the test score

feedback condition did not intend or expend more effort or perform better on the PISA math items than students given the standard test instructions (i.e., "control" condition).

The type of feedback promised also does not seem to affect examinee motivation or performance. V. L. Wise (2004) investigated whether examinees expended more motivation depending on the kind of feedback promised. Four conditions were created which varied the instructions given before a low-stakes exam on information literacy: (a) norm-referenced, (b) criterion-referenced, (c) choice of feedback, and (d) no feedback. In the norm-referenced condition, participants were told that their individual test score would be compared to the mean test score of other students at the university. In the criterion-referenced condition, participants were instructed that their individual score was to be compared to a predetermined standard. In the choice condition, participants were given the option of either norm-referenced, criterion-referenced, or no feedback at all. Self-reported motivation as measured by a total score on the Student Opinion Scale (SOS; Sundre & Moore, 2002) did not differ among the various feedback conditions. However, within the choice condition, participants who chose either norm- or criterion-referenced reported higher examinee motivation that those who chose no feedback. Students in all experimental and non-experimental (i.e., choice) conditions performed similarly on the information literacy test.

Snyder (2012) explored whether giving feedback about performance after the first administration of a low-stakes test would increase self-reported perceived importance of the test and therefore increase examinee effort for the second administration of the test. Seventh and eighth grade students were given the 4Sight, an achievement questionnaire comprised of constructed and selected-response items in math and reading. This test is

typically used for teacher feedback on student progress and to predict performance on a statewide accountability exam. All students completed the achievement questionnaire and the Student Opinion Scale (SOS; Sundre & Moore, 2002) measuring effort and importance at two time-points during the school year. Students were randomly assigned to either an experimental or control condition. In the experimental condition, students received coaching regarding the content of the 4Sight tests, its importance, and feedback on their pretest scores. The control group did not receive coaching. Overall, there was no effect of the intervention on either effort or importance. That is, both groups reported the same mean levels of test importance and test-taking effort across time. For both groups, self-reported importance means did not decline over time; however, effort did decline over time.

Reporting test-score feedback on a low-stakes exam has not shown sufficient promise in increasing overall examinee motivation. Although students report that they are interested in receiving feedback on their low-stakes test performance (Kiplinger & Linn, 1995), very few students seem to actually seek out this information when offered. Socha, Swain, and Sundre (2013) examined whether the promise of feedback on a low-stakes assessment affected choice of the type of feedback. Students were randomly assigned to testing conditions that varied by test instructions: one group was told that individual interpretive feedback would be sent to the student later in the semester. The other group did not receive these instructions. Both groups received an email later in the semester indicating that individual feedback was available for their review[1]. Feedback was offered

[1] Readers may recall the study by Finney, Sundre, Swain, and Williams (2015) where students were promised score feedback on their low-stakes institutional assessment. This study further explored the Feedback condition.

in two ways: (a) their individual score compared to peers or "norm-referenced" or (b) their individual score compared to faculty standards or "criterion-referenced." Students were given the choice as to which kind of feedback to obtain. Of the students who were told they would receive feedback, only 30.4% actually sought this information. A similar number of students who were not promised feedback actually sought feedback of any type (31.3%). Across both groups, students who sought feedback were primarily interested in their scores compared to other students with about 65% of students in each group choosing to receive norm-referenced feedback. If incentivizing assessment test performance (i.e., consequences, score feedback) does not improve motivation, perhaps practitioners should consider the cost to examinees associated with completing these low-stakes exams.

Considering Examinee Burden

Several strategies to increase examinee motivation have been reviewed. Informing students of test score use, providing monetary incentives, and test score feedback have not shown sufficient promise. Tying low-stakes test performance to grades does appear to motivate students to perform well. In essence this practice changes the nature of the exam to high-stakes. As discussed, this policy may be difficult for assessment practitioners to implement, and motivation may be dependent on the amount of credit awarded. If assessment practitioners are not able to tie assessment test performance to grades, what then is a possible solution? Given the lack of a consistently effective motivational strategy in the literature compels one to reflect more deeply on examinee motivation theory.

A common thread in both EV theory and the DC model is the concept of *cost* or *fatigue*, which is hypothesized to decrease motivation. In fact, *mental taxation* is described as a predictor of both effort capacity and resource demands in the DC model, which are then predictors of effort on a given item. Perhaps reducing testing burden would move the needle of examinee effort where previous attempts have not shown sufficient promise (S. L. Wise & DeMars, 2005). Although other constructs reviewed previously have been theorized to *positively* relate to effort (i.e., test score use, academic citizenship, grades, money, feedback), there are theorized aspects of the testing context that *negatively* relate with expended effort on a test. Cost has been shown to negatively affect behavior even when value is high. Battle and Wigfield (2003) found that perceived value of attending graduate school positively related with intention to attend but the psychological cost of attending graduate school negatively related to their intentions. In this way, high value may increase an individual's chance of engagement but high cost can certainly reduce these chances.

In a testing context, students may perceive the task to be important (i.e., high value) but may also perceive the arduousness or length of the test to be a high cost. By reducing these hindrances (i.e., test length), test-taking effort may increase simply because the test is more enjoyable or less of a burden (Lau, Swerdzewski, Jones, Anderson, & Markle, 2009). Several researchers have discussed the possibility of observing an increase in examinee motivation via reducing the cost associated with an exam. Test length may be a determinant of examinee test-taking effort in higher education assessment contexts given that completing a demanding low-stakes test for many hours can be perceived as "costly" by examinees (Graham, 2012). Perhaps some

authors' notions regarding shorter tests (S. L. Wise, 2006) or a matrix sampling technique (DeMars, 2007) could allay the effect of reduced test-taking effort within a test. Moreover, administering shorter tests via a PMD may *increase* the validity of test scores by reducing examinee burden or cost (T. D. Little et al., 2014). That is, PMDs may lead to higher data quality as a result of decreased examinee burden when compared to administering a full-form of an instrument (T. D. Little & Rhemtulla, 2013). Assessment practitioners can control the burden on any one examinee through a PMD reviewed in Chapter 1. Although no studies have attempted to reduce examinee fatigue, some studies have discussed examinee fatigue broadly within the low-stakes testing context.

Examinee Fatigue during Low-Stakes Exams

Some researchers have explored examinee motivation over the course of a low-stakes testing session in higher education. Barry, Horst, Finney, Brown, and Kopp (2010) employed mixture modeling to assess whether classes of examinees emerged that changed in levels of test-taking effort as the testing session progressed. Self-reported effort was measured by the Student Opinion Scale (SOS; Sundre & Moore, 2002) which was administered following each test. The testing session consisted of two non-cognitive exams followed by a cognitive exam followed by two more non-cognitive exams. Self-reported test-taking effort data were collected following each test: examinees responded to five test-taking effort items on a 5-point Likert scale from 1 "Strongly Disagree" to 5 "Strongly Agree." Barry and colleagues reported three latent classes of examinees with distinctive patterns of change in test-taking effort. Class 1 (about 8% of the sample) reported, on average, similar effort on the non-cognitive tests but lower effort with the cognitive test. This class had relatively high test-taking effort scores with mean effort

scores > 4.50 on a 5-point scale for most tests. Class 2 (about 71% of the sample), on average, followed the same pattern as Class 1 with less expended effort on the cognitive test than the non-cognitive tests, but this class also reported consistently lower effort on every test than Class 1 (about a 1-point lower mean effort score on each test). Finally, Class 3 (about 21% of the sample) reported consistent test-taking effort across all five tests, on average, with each mean effort score around a 4.

Given no evidence of a systematic reduction in test-taking effort as the testing session progressed, Barry et al. (2010) denounced the notion of examinee fatigue. Since evidence of a steady decline in self-reported test-taking effort as the session progressed was not observed, the authors posit that concern about test placement within an educational assessment testing session is not warranted. However, other researchers have found that rapid guessing behavior, an indicator of low examinee motivation, increased while performance decreased for students given the same test later in a series of tests (DeMars, 2007). Moreover, most students reported lower effort for the cognitive exam in the Barry and colleagues' (2010) study, which may be due to the difficulty of that exam. Responding to self-report items on non-cognitive assessments is not as difficult as a test of scientific reasoning skills; hence the lower reported effort for the latter. Thus, cost of task engagement seems to be a factor that contributes to fatigue and reduces performance.

Although some students may have given sustained effort across multiple tests, this effect does not necessarily mean that these same students gave sufficient effort on all items within a single test. Differential effort on some items may result in item position effects, which are differences in item parameters due to the serial position of the item. The DC model asserts that motivation may change as an examinee proceeds through a

test and that effort expended on any given item may not be the same as the effort given on previous items (S. L. Wise & Smith, 2011). Item position effects have been documented in the item response theory (IRT) literature. This means that the qualities of items are affected by placement within a test above and beyond other factors that influence an item's psychometric quality. Examples of examinee fatigue effects at the item level include rapid guessing, omissions, and changes in item difficulty and discrimination parameter estimates.

Rapid guessing and omissions. Rapid guessing appears to be particularly pronounced in low-stakes testing contexts and may be explained by low motivation. Setzer, S. L. Wise, van den Heuvel, and Ling (2013) collected data on the Major Field Test in Business, a low-stakes exam. The researchers calculated response-time fidelity (RTF) scores for each item (S. L. Wise, 2006). RTF is the proportion of examinees in the sample who are engaging in *solution behavior* as opposed to *rapid guessing behavior*. Solution behavior refers to whether the examinee spent enough time on task to read and respond thoughtfully to an item, whereas rapid guessing behavior is the opposite of solution behavior. Both of these indices are collected during computer-based exams where progress on every item can be monitored. The authors noted that rapid guessing behavior on an item was related to the item's difficulty (p-value), position, and length. That is, the more difficult an item, the farther it was along in a test, and the more reading involved, the less effort examinees were likely to expend. Other authors have found very similar findings in low-stakes contexts. S. L. Wise, Pastor, and Kong (2009) found that students were more likely to rapidly respond to items with more text, those appearing later in the test, and those with greater numbers of response options during a low-stakes

exam. Models have been proposed to capture rapid guessing behavior (Cao & Stokes, 2008) as well as to identify the point at which examinees switch from solution behavior to rapid guessing behavior based on their pattern of correct and incorrect responses on a test (Yamamoto, 1995).

When examinee motivation is low, some examinees may choose to omit an item response rather than rapidly guess. Omitted item responses may occur within a test for various reasons. For some examinees, the test may be *speeded* or require more answers than time allows. Others may not be motivated to give a response. As part of a series of studies, the National Center for Education Statistics investigated reasons why students were omitting items on the NAEP exam. Some students reported that they did not know the answer, thought the question would take too much time, or cited motivational issues with long, complicated items (Jakwerth et al., 1999). Motivational issues were more often cited as a reason for not responding to constructed-response items over multiple-choice items. From an assessment practitioner's perspective, omitted responses due to a lack of motivation can jeopardize inferences about student learning, particularly for constructed-response items.

Sometimes examinees do not omit answers when they run out of time; instead, they provide rapid, random responses to items further along in the test due to a steady decline in effort. S. L. Wise (2006) noted that even when examinees are given sufficient time, rapid guessing behavior may still occur because of low test-taking effort. Guessing and omitting item responses can attenuate ability estimates (De Ayala, Plake, & Impara, 2001); therefore, the connection between test length and psychometric data quality should be apparent. The effect of item position on item parameters (i.e., difficulty and

discrimination) is reviewed next to examine possible benefits of shorter tests on assessment data quality.

Item difficulty. Evidence of item position effects on item difficulty appears to be mixed. In low-stakes contexts, some evidence has suggested that item difficulty *increases* as the item moves in position further in the test. Debeer and Janssen (2013) analyzed PISA data that was administered in a rotated block design. That is, the same block of items were shifted into different positions and each block contained science, math, and reading items. Thirteen forms were created so that each block was given in first, second, third, and fourth in order. This design allowed modeling of the position effect of each item by subject. Using a 2PL IRT model, the authors found that as the same item moved farther down in the test position, the difficulty of the item increased. This effect was largest for reading items in which IRT difficulties increased by .240 per cluster position shift. Additionally, the authors reported that the correlation between latent ability and the item position dimension (i.e., change in item difficulty) was -.257 for science, -.531 for reading, and -.357 for math. This indicates that an increase in item difficulty was less pronounced for high ability examinees. Given the low-stakes nature of the PISA test, the authors noted that lack of examinee effort could explain this item position effect.

In addition to the PISA study, the same authors analyzed data from a listening comprehension exam given to eighth grade students in Belgium (Debeer & Janssen, 2013). This test was split into three item sets (i.e., A, B, C) which were assembled into four forms (AB, BA, BC, and CB) that were randomly assigned to a sample of examinees. Using a 2PL IRT model with mean and variance of the latent ability fixed to 0 and 1, respectively, differential item functioning (DIF) was employed to examine

changes in item difficulty between the forms. The authors noted that as item position moved further along in the test, item difficulty increased. The nature of this trend was tested and was found to be linear and resulted in a fairly large and positive correlation (r = .71) between item position and item difficulty.

In high-stakes contexts, item position effects on item difficulty appear to be reversed. Kingston and Dorans (1984) analyzed Graduate Record Examination (GRE) General Test data where two different forms were administered. In Form A, four sets of operational items were given followed by a fifth set of items that contained operational items from Form B. Form B also contained four sets of operational items followed by a fifth set of items from Form A. The fifth sets of both forms varied which operational items were borrowed from the other forms. All items from the four original sets of Form A were given again in a later position on Form B and vice-versa. In this way, the same items were given in two different positions, allowing the researchers to assess item position effects. IRT item difficulty (b-parameters) were found to *decrease* (i.e., items were easier) for items appearing later in the test, particularly for items with complicated directions. The authors refer to this decrease in item difficulty as a *practice effect* as the authors hypothesized that examinees needed to become familiar with the directions of these item types by completing a few of them before they became easier. These item types have since been removed from the GRE General Test due to this effect (Swinton, Wild, & Wallmark, 1983) indicating that investigating position effects could have major implications for testing programs.

Others have found no effect on item difficulty. L. S. Rubin and Mott (1984) examined data collected from a routine assessment of the Virginia Minimum

Competency Reading Test. However, the stakes of this test were not described but can be assumed to be high. Using a Rasch model, the researchers found no effect of item position change on item difficulty. Classical Test Theory (CTT) difficulty parameters (p-values) were also examined. These values also did not appear to change greatly as the items moved position. Although effects of item position on item difficulty have resulted in some mixed findings, when the test context is known to be low-stakes to students, item difficulty increases with item position. Although more research has concerned changes in item difficulty due to item position change, a few researchers have examined changes in item discrimination.

Item discrimination. Meyers, Murphy, Goodman, and Turhan (2012) investigated the impact of item difficulty and item discrimination in two statewide assessment testing programs. Using a 3PL IRT model, the researchers found that, on average, both item difficulty and item discrimination *increased* when the same item appeared later in the test. That is, items became more difficult but were also more discriminating when they appeared later in the test. These effects appeared to be larger when the item position change was greater.

Outside of educational assessment, authors have investigated item position effects on item discrimination for some time. Knowles (1988) analyzed data from four different self-report personality measures. CTT item discrimination parameters (item-total correlations) were computed for each item at each position. These item discrimination parameters increased as the same item moved position further along in the test. Steinberg (1994) found evidence countering the presence of this effect. Using a graded-response IRT model on data from an anxiety measure, an omnibus test of differences in item

discrimination parameters was significant. However, when examining the differences between each item, this effect was only significant for one item suggesting a specific context effect. Of these studies, only one concerned an assessment testing program an each suggest disparate conclusions. PMDs change the item position within each form; therefore, differences in item parameters should be explored to assess the psychometric quality of items administered in a PMD.

The results of these studies suggest that item position may impact item parameters. Notably, the most consistent effects of item position appear in low-stakes testing contexts. The framework of examinee motivation, and the tenets of examinee fatigue, may explain these observed effects. However, more research is clearly needed to explore whether these item position effects are a result of low examinee motivation and, if so, how to mitigate their influence. If item position effects are observed during low-stakes assessments (e.g., greater item difficulty at the end of a long test), perhaps PMDs could mitigate these effects. That is, PMDs could address examinee cost and fatigue while increasing data collection efficiency. The effect of PMDs on item-level data quality will be explored in this dissertation. With planned missing data, assessment practitioners may wonder what missing data handling technique is most appropriate to obtain accurate item parameters, group-level estimates, and standard errors. A thorough review of traditional missing data handling techniques, and their limitations, should point to the suggested use of a modern technique to recover data missing by design.

Traditional Missing Data Handling Techniques

Several techniques are available to handle missing data, including data missing by design. The goal of this section is to provide an overview of traditional and modern

missing data handling techniques when data are missing by design. When using a PMD, the missing data mechanism is under the researcher's control. Recall that PMD data are known to be MCAR by design even though there may be other missing data mechanisms in the dataset. Fortunately, most missing data handling techniques assume that data are MCAR where the propensity for missingness is unrelated to any variable; however, some techniques allow for a MAR mechanism where the propensity for missingness is related to some other variable in the analysis model (D. B. Rubin, 1976). Severe limitations exist with most traditional techniques even when data are assumed to be MCAR. These traditional techniques are problematic in that their implementation produces biased parameter estimates and standard errors. The exact biases of each technique when used with data collected via a PMD are summarized in Table 2. Several approaches for dealing with missing data are available and will be described in more detail in the forthcoming sections.

Deletion techniques. One approach to handling missing data is simply to delete cases where missing data exists. Listwise deletion (LD) involves deleting a case that has missing data on any variable to be analyzed and results in a complete dataset. In this way, all analyses are based on the same sample of participants. Pairwise deletion (PD) is less aggressive in that only cases that have complete data *for each analysis* are included. This means that if several variables within a dataset have missing data, the sample size may vary by analysis. For example, a correlation between X and Y may be computed using a different sample than a correlation between Y and Z. In a review of several educational and psychological journals, 96% of the studies that discussed handling missing data used LD, PD, or both, making these approaches the most popular (Peugh & Enders, 2004).

Table 2

Biases of Missing Data Techniques When Using Planned Missingness Designs Compared to a Full-Form Design

Method	Assumed Missingness Mechanism	Test/Item Mean	Item Correlation	Standard Error of the Test/Item Mean	Standard Error of Item Correlation	Comments on Method with PMD Data
Listwise Deletion (LD)	MCAR	Unbiased	Unbiased	Biased Upward	Biased Upward	Not advisable as all cases would be deleted (unless one group completed all items)
Pairwise Deletion (PD)	MCAR	Unbiased	May Exceed Bounds [a]	Biased Upward or Downward [b]	Biased Upward or Downward [b]	Not advisable when estimating parameters that use one sample size
Arithmetic Mean Imputation (AMI)	None	Unbiased	Biased Downward	Biased Downward	Biased Downward	Never advisable
Regression Imputation (RI)	MCAR	Unbiased	Biased Upward	Biased Downward	Biased Downward	Not advisable as superior techniques exist
Stochastic Regression Imputation (SRI)	MCAR, MAR	Unbiased	Unbiased	Biased Downward	Biased Downward	Generally appropriate when estimating means however superior techniques exist
Multiple Imputation (MI)	MCAR, MAR	Unbiased	Unbiased	Unbiased	Unbiased	Recommended in all scenarios except for MNAR data
Full-Information Maximum Likelihood (FIML)	MCAR, MAR	Unbiased	Unbiased	Unbiased	Unbiased	Recommended in all scenarios except for MNAR data

Note. All missing data handling techniques consider biases when using data that is *only* missing by design (MCAR). Additional missing data mechanisms may exist within a single data set (i.e., MAR) and may vary by analysis. MI and FIML perform similarly with MAR data. Cells shaded in gray present limitations of each technique compared to a full-form design.

[a] Correlations may exceed bounds when covariances are computed using one set of complete cases between variables X and Y but a different set of complete cases to calculate variances for X and Y.

[b] If PD is used to estimate a covariance matrix for use in a regression, for example, each covariance may be computed from different samples of varying sizes. Therefore, there is no one sample size to use for the standard error formula. If an average sample size is used, some standard errors will be biased upward while others are based downward (R. J. A. Little, 1992).

LD and PD have several limitations when used to estimate parameters and standard errors on data collected via a PMD. First, both techniques assume the missing data mechanism is MCAR. If the MCAR assumption holds, both LD and PD generally result in unbiased mean and correlation parameter estimates (Arbuckle, 1996). When the missing data mechanism is MAR, both LD and PD result in biased parameter estimates. Moreover, these methods discard potentially important information and reduce the statistical power of each analysis when including a variable that has missing data. Most troublesome, using LD on PMD data where no students complete all items would result in deleting all cases. Finally, PD presents a unique problem when estimating item correlations. Software packages may compute a covariance between two items using the cases that have data on those two items only. In order to calculate a Pearson correlation, this covariance is divided by the square root of the product of the standard deviations for each item. The problem arises when these standard deviations are calculated using all available cases for each item instead of the same sample used to compute the covariance. When this approach is used, some correlations may exceed 1 (R. J. A. Little, 1992). PD also contributes to the problem of *non-positive definite matrices*, which are correlation or covariance matrices that are mathematically impossible (Enders, 2010). Non-positive definite matrices can result in estimation issues when fitting multivariate models such as multiple regression or structural equation models. Due to these issues, PD is not recommended to handle missing data.

Single imputation techniques. Rather than discarding data, several techniques exist to *impute* or replace a missing value with a plausible value. These techniques are referred to as single imputation techniques because they replace a missing value with a

single value. The first is arithmetic mean imputation (AMI), or mean substitution in which the mean of the variable computed using complete cases, is used to replace missing values for that variable. For example, if 10% of cases in a dataset have a missing value for variable X, the mean of variable X computed from the 90% of the cases with complete data is substituted for the missing values. AMI inherently reduces variability in the dataset and can severely bias variances and subsequently standard deviations downward (Enders, 2010). This technique also attenuates or reduces the magnitude of correlations from their true values and overestimates standard errors (Schlomer et al., 2010). Moreover, AMI results in biased parameters and standard errors even when the missing data mechanism is MCAR. For these reasons, AMI is never an acceptable methodology to handle missing data.

In an attempt to use the relationships among variables in a dataset to improve imputation, regression imputation (RI) was developed (Buck, 1960). RI first builds a regression equation between the variable with missing data (Y) and other variables in the dataset with available cases (X). This equation is then used to predict missing values on the variable with missing data. While RI is an improvement from AMI, there are still limitations to it use. For example, all predicted values fall perfectly on the regression line. If one IV was used in the regression model, this assumes that the IV and the DV with missing data are perfectly related, which is highly unlikely. Consequently, RI overestimates correlations, regression coefficients, R^2 values, and variability (Graham, Cumsille, & Elek-Fisk, 2003). Whereas means are unbiased under MCAR or MAR (Schlomer et al., 2010), RI is also not recommended due to biases in measures of association.

Stochastic regression imputation (SRI) attempts to overcome the upward bias of measures of association from RI by restoring variability in the estimated values. Just like RI, SRI predicts missing values using a regression equation but adds a random (stochastic) error term.

$$Y_i^* = \left[\widehat{\beta_0} + \widehat{\beta_1}(X_i)\right] + z_i \quad (1)$$

In Equation 1, Y_i^* is the predicted value for person *i* using their observed data on variable X. The random term (z_i) is normally distributed with a mean equal to zero and a variance equal to the residual variance from the regression equation. Therefore, means are unbiased since the mean of this term is zero and variances are unbiased under a MAR mechanism (Enders, 2010). However, standard errors tend to be biased downward resulting in higher Type I error rates (Baraldi & Enders, 2010). Due to this limitation, methodologists recommend employing one of two modern techniques that produce more accurate standard errors than SRI.

Multiple Imputation

Multiple imputation (MI) is one missing data handling technique that is considered state-of-the-art (Schafer & Graham, 2002). MI functions similarly to the SRI technique where only one dataset is generated with complete data. Like SRI, MI fills in missing values with predicted responses. However, MI produces multiple datasets with slightly different predicted responses for each missing value in each dataset (Schafer, 1997). Overall, the process of MI involves three phases: (a) imputation, (b) analysis, and (c) pooling. Each stage is described separately.

The EM algorithm. The first step is referred to as the *imputation phase* where a specified number of datasets ($m > 1$) are generated. The imputation phase first begins with maximum likelihood estimates of the mean vector ($\widehat{\boldsymbol{\mu}_1}$) and covariance matrix ($\widehat{\boldsymbol{\Sigma}_1}$) of the dataset. The expectation-maximization (EM) algorithm is commonly employed to provide these initial estimates (Dempster, Laird, & Rubin, 1977). The EM algorithm iterates between the E-step (expectation) and the M-step (maximization) until converging on a single set of mean vector and covariance matrix estimates. Conceptually, the first E-step obtains a mean vector and covariance matrix using observed data to "fill-in" missing data via several stochastic regression equations (Enders, 2010). Actually, no values are imputed but information is borrowed from the complete data to provide the *sufficient statistics* for estimating a mean vector and covariance matrix: the sum of scores, sum of squares, and sum of cross-products. These sufficient statistics are then used to compute means, variances, and covariances for all variables in the M-step (Enders, 2010). This new mean vector and covariance matrix is then used in the next E-step to estimate new regression equations for the sufficient statistics. This process iterates between E- and M-steps until the log-likelihood increase between M-steps falls below a set criterion.

Using the final estimated mean vector ($\widehat{\boldsymbol{\mu}_1}$) and covariance matrix ($\widehat{\boldsymbol{\Sigma}_1}$) from the EM algorithm, a Markov chain Monte Carlo (MCMC) or data augmentation algorithm begins. Using a Bayesian framework, the goal of this algorithm is to find a stable posterior distribution of the parameters from which *m* datasets are drawn. This MCMC procedure assumes multivariate normality but has been shown to be robust to violations of this assumption (Graham & Schafer, 1999). This algorithm iterates between an imputation step (I-step) and a posterior step (P-step). Using the estimates from the EM

algorithm, stochastic regression equations are built to predict missing values and create a complete dataset with some inter-individual variability built in with the random residual (see Equation 1). This step is referred to as the I-step or the imputation step. Instead of basing every imputed dataset on the same mean vector and covariance matrix, the parameter estimates of the sample are perturbed with a random residual term. Subsequently, a new mean vector ($\widehat{\boldsymbol{\mu}_2}$) and covariance matrix ($\widehat{\boldsymbol{\Sigma}_2}$) is then estimated in this P-step as this process mimics drawing new estimates from a posterior (or sampling) distribution in a Bayesian framework. Another dataset is imputed from this new mean vector and covariance matrix via stochastic regression and the I-step and P-step process iterates until specified to end (i.e., *m* datasets are created). Datasets from immediate P-steps are typically highly correlated. The researcher must specify the number of iterations between imputations so that each dataset is independent. Moreover, convergence of the data augmentation algorithm is trickier than EM since convergence here is defined as the stability of the posterior distribution.

Assessing imputation model convergence. There exist several ways to assess convergence of the data augmentation algorithm. These include visual inspection of *time-series* and *autocorrelation function plots* and the *worst linear function* (Enders, 2010). For each parameter in the mean vector and covariance matrix, time-series plots depict the simulated parameters at each P-step iteration. Parameter estimates that jump around a single point in a seemingly random fashion signal a stable posterior distribution. When parameters veer off in one direction for several cycles, the parameter's posterior distribution is not stable. For example, if a systematic trend occurs for 50 iterations then at least 50 iterations should separate random draws. Autocorrelation function plots depict

the dependency between P-step parameter estimates. Several Pearson correlations between a parameter's simulated values and itself at lag-*k* is computed. The researcher examines the plots to determine the number of iterations (lags) needed before the simulated parameters are independent. Schafer (1997) proposed using the worst linear function (WLF) as a summarizing indicator of convergence. The WLF is a weighted sum of the parameters that converged the slowest (i.e., largest change at final iteration). The WLF values at each P-step are displayed on a time-series plot to indicate systematic trends in data convergence speed. A similar visual analysis of systematic patterns could indicate the number of iterations between each data set.

Convergence problems in the data augmentation algorithm can arise when the rate of missing data is very high or the number of variables exceeds the number of cases. Occasionally, stable posterior distributions cannot be reached by the algorithm within a default number of iterations. This may occur with high multicollinearity among the variables in the imputation model. When practitioners encounter convergence issues, there are a few options. By default, the imputation software typically specifies a non-informative prior. In this way, the data alone dictate $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$. When convergence problems arise, specifying *prior information* or using *a ridge prior* may stabilize the data augmentation algorithm. Prior information involves providing the algorithm with reasonable estimates of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ that can be combined with the data to produce posterior distributions. The downside of prior information is that the resulting posterior distributions are more similar to the sample that produced the prior information. Another option is to specifying a ridge prior distribution. Conceptually, a ridge prior adds very few imaginary cases with complete data to the dataset. However, the covariance matrix of

these imaginary cases has covariances equal to 0. Thus, relationships between variables in the imputation model are biased slightly downward. This practice is intuitively helpful when convergence problems are suspected due to high multicollinearity.

The number of datasets to be generated must be specified by the researcher or assessment practitioner. Methodologists have conducted simulation studies to determine how many imputed datasets are sufficient for accurate standard errors. These early studies recommended three to five imputed datasets (Schafer & Olsen, 1998); however, subsequent simulations suggested that more imputations may be necessary. Statistical power has been shown to improve when the number of imputations increases beyond 10. Unless the fraction of missing information, defined later, is greater than .50, there is a diminishing return when specifying more than 20 imputed datasets (Graham, Olchowski, & Gilreath, 2007). Described in more detail later in this chapter, FMI captures the proportion of sampling variability that is due to missing data and is usually *not* analogous to the percent of missing data. In some PMDs on a single test, data can be missing on 50% of the test across the sample (i.e., BIB design). Although a good imputation model (i.e., correlated variables) will likely not result in FMI values approaching .50, 40 imputations is recommended. Practitioners should note that, although more imputations is preferred to fewer, having as many as 100 imputed datasets may result in slower computation times when running complicated models.

After *m* datasets have been imputed, the second step of MI begins. In the *analysis phase*, the desired analysis specified by the researcher is performed on each multiply-imputed dataset separately. For example, consider an assessment practitioner who collected data via a PMD on a 50-item test of information literacy. She wants to compute

a total test mean for the sample, so she runs the imputation step of an MI procedure and generates 20 complete datasets. She sums each participant's imputed and observed item responses within each imputed dataset. She then computes a mean of these total scores for the entire sample within each of the 20 datasets resulting in 20 means. At this point, she has completed the analysis phase. The results of each analysis is combined in the *pooling phase* to produce a single set of results using D. B. Rubin's (1987) rules.

$$\overline{\mathrm{X}} = \frac{1}{m}\sum_{t=1}^{m}\widehat{\mathrm{X}_t} \tag{2}$$

Usually, parameter estimates are simply the arithmetic mean of the parameter estimates within each dataset ($\widehat{\mathrm{X}_t}$) across the *m* datasets. In the total test mean example, the researcher would use Equation 2 to take the arithmetic mean of her 20 means. Pooling standard errors for these parameter estimates requires a few more steps. For the total test mean example, our researcher would use Equations 3 through 5 to compute an accurate standard error for her single total test score mean.

$$\mathrm{V_W} = \frac{1}{m}\sum_{t=1}^{m} SE_t^2 \tag{3}$$

With 20 means, there are 20 standard errors for these means. Equation 3 is simply the mean of the squared standard errors (i.e., sampling variance) produced *within* each imputation.

$$\mathrm{V_B} = \frac{1}{m-1}\sum_{t=1}^{m}(\widehat{\mathrm{X}_t} - \overline{\mathrm{X}})^2 \tag{4}$$

Equation 4 incorporates uncertainty as a result of imputing missing values. Recall that, like SRI, MI incorporates a random error term when predicting missing values.

Therefore, the predicted values ($\widehat{X_t}$) for each missing data point fluctuate between each dataset somewhat. The variance of these predicted values around their mean captures the sampling variability *between* each imputation. This between-imputation variance (V_B) component is then combined with the within-imputation variance (V_W) to form total variance (V_T).

$$V_T = V_W + V_B + \frac{V_B}{m} \tag{5}$$

Equation 5 combines within- and between-imputation variance into total variance (V_T). The $\frac{V_B}{m}$ term serves as a correction factor due to the use of a finite number of imputations (Enders, 2010). With more imputations, this term becomes smaller. For example, if our researcher wants to obtain a standard error estimate for her single total test score mean, she would take the square root of the V_T term to transform from the sampling variance metric to the standard error metric.

Within-imputation (V_W) and between-imputation (V_B) variance can be combined to form two other indices, which can assist practitioners in assessing the trustworthiness of the MI estimates. These indices are degrees of freedom (v) and fraction of missing information (FMI). For each parameter estimate, v is computed using Equation 6 (Enders, 2010).

$$v = (m - 1)\left[1 + \frac{V_W}{V_B + \frac{V_B}{m}}\right]^2 \tag{6}$$

Despite its name, v has nothing to do with sample size. This index may range from $v = m - 1$ to $v = \infty$ with higher degrees of freedom with more imputations or when fraction of missing information is low (Enders, 2010). Practitioners may conceptually interpret

this index as a measure of the stability of MI estimates with higher values more desirable (Graham, 2012). That is, when v approaches ∞, all variation in the parameter estimate is sampling variation and not due to missing data (van Buuren, 2012).

Barnard and Rubin (1999) noted that v could exceed complete-data degrees of freedom, and simulations suggested v is inappropriate in small samples. They suggest an adjusted degrees of freedom (v_1),

$$v_1 = \left(\frac{1}{v} - \frac{1}{\tilde{v}}\right)^{-1} \tag{7}$$

where

$$\tilde{v} = (1 - \text{FMI})\left(\frac{edf + 1}{edf + 3}\right) edf \tag{8}$$

and edf is the complete-data degrees of freedom. FMI is defined in Equation 9 and described below. The adjusted degrees of freedom now increases with sample size and does not exceed degrees of freedom if the data were complete (edf). The v_1 estimate of degrees of freedom is used to construct confidence intervals around a parameter estimate and is preferred over v no matter the sample size. Fortunately, v_1 and the appropriate confidence intervals are available from MI software.

FMI is another way to assess stability of MI estimates as it captures the influence of missing data on sampling variance for each parameter (Enders, 2010).

$$\text{FMI} = \frac{\text{V}_\text{B} + \frac{\text{V}_\text{B}}{m} + 2\text{V}_\text{W}/(v + 3)}{\text{V}_\text{T}} \tag{9}$$

Looking at Equation 9, FMI defines the proportion of variation added by missing data to the squared standard error (V_T). When variables in the imputation model do well

recovering information in the missing data, FMI will be small. At its upper limit, FMI is equal to the percentage of missing data, but that will only occur when the variables in the imputation model are not correlated with the missing values. In the case of PMDs, FMI values close to zero and far below the percentage of planned missing data are preferred.

Mentioned previously, parameters are *usually* the arithmetic mean of the parameter estimate across imputations. This rule is dependent on the exact parameter of interest. Parameter pooling is improved when the scale of the parameter estimates are close to normally distributed (van Buuren, 2012). Therefore, some parameters should be transformed prior to pooling. For example, correlations should be transformed to *z*-score metric, pooled, and then the final estimate may be transformed back to correlation metric using Fisher's *r* to *z* formula (Thorndike, 2007). Similarly, R^2 values should be transformed using Fisher's *z* on root. For the interested reader, several other less common parameter estimate transformations are described by van Buuren (2012, p. 156).

The use of auxiliary variables. MI assumes that the missingness mechanism is MAR (Schafer & Olsen, 1998). The tenability of this assumption can be improved by the inclusion of *auxiliary variables* (Baraldi & Enders, 2010). Moreover, the use of auxiliary variables in MI tends to improve the power of significance tests even when data are MCAR (p. 21). Auxiliary variables assist in the imputation of the missing values as they correlate with the variable that is missing (Y) or help explain the missingness mechanism (R). An auxiliary variable has been suggested as "good" if $r > .40$ between the auxiliary and the missing variable (Collins, Schafer, & Kam, 2001). Collins and colleagues suggest that practitioners should err on including more auxiliary variable than fewer; however, this "inclusive" strategy may bias estimates when the missingness mechanism is not

MCAR (Thoemmes & Rose, 2014) or with small sample sizes and low correlations between the auxiliary variables and the missing variable (Hardt, Herke, & Leonhart, 2012).

In higher education assessment, most variables of interest will be correlated with each other to some extent. Within a single test, individual items are likely to correlate with each other at least moderately. Therefore, in a PMD, available items may be used as auxiliary variables to assist MI in providing accurate estimates for missing items. In a three-form design, for example, items or scales in the X-set are included on all forms. If these items or scales are at least moderately correlated with the missing data, then they should be included as auxiliary variables. Moreover, imputation at the item-level has been suggested as superior to scale-level imputation (Gottschall, West, & Enders, 2012). That is, PMDs that exclude items from a single test or a battery of tests melds well with suggested practice to impute missing items rather than scale scores.

Rounding imputed values. As a final note, when imputing data that are not continuous (i.e., ordinal) using the usual regression procedure, imputed data will have decimals. Practitioners have traditionally been advised to round the imputed values to the nearest sensible integer (Rubin, 1987). For example, a predicted value of 3.579 using data collected on a 5-point Likert type scale would be rounded to 4. Rounding predicted values of variables with slight departures from normality (e.g., when ordinal variables can take on many categories) may result in negligible bias (Schafer, 1997). However, some binary variables (e.g., gender) may produce biased results with this approach. In the case of binary variables, an approach known as "naïve" rounding uses a cut-off of .5 to round imputed data to values of 0 or 1. Naïve rounding produces bias in means and

correlations whereas the analysis of unrounded binary data resulted in means (i.e., proportions) and correlations very close to true values (Horton, Lipsitz, & Parzen, 2003; Yucel & Zaslavsky, 2003).

Allison (2005) examined several methods for imputing binary data using the PROC MI procedure in SAS: (a) listwise deletion (no imputation), (b) linear imputation without rounding, (c) linear imputation with rounding, (d) logistic regression imputation, and (e) discriminant function imputation. The accuracy of mean estimates (i.e., proportions) and regression coefficients of the binary variable/dummy code were compared across four levels of true means: .50, .20, .05, and .01 and two missing data mechanisms: MCAR and MAR with 50% missing data. Sample size was fixed to 500 in all conditions. All methods except linear imputation with rounding performed equally well when estimating the regression coefficient for a dummy variable. In general, linear imputation with rounding produced the most bias compared to all other methods. When estimating means, listwise deletion resulted in larger standard errors than other methods. In the conditions when means are small (i.e., < .20), logistic and discriminant function approaches were superior to linear imputation without rounding but were about equal in other conditions. However, logistic and discriminant function methods are only applicable when data have monotone missingness. Missing data patterns may be either monotone or arbitrary. Figure 2a and 2b display these missing data patterns graphically. If a researcher is interested in using logistic or discriminant analysis for his or her imputation model, the MCMC method may be used to first impute *some* data to result in a monotone missingness pattern. Then the researcher may change the imputation model.

Case	X1	X2	X3	X4
1	O	O	O	O
2	O	O	O	M
3	O	O	M	M
4	O	M	M	M
5	M	M	M	M

Figure 2a. A monotone missingness pattern. "O" indicates the data point is observed; "M" indicates the data point is missing. Note that Case 1 has complete data on all variables whereas Case 3 is missing data on the last two variables: X3 and X4. Variable order matters in the definition of monotone missingness. If Case 3 were missing data on variable X1, the missingness pattern would be arbitrary, not monotone.

Case	X1	X2	X3	X4
1	O	O	M	O
2	M	O	O	M
3	M	M	O	M
4	O	O	M	O
5	M	O	O	O

Figure 2b. An arbitrary missingness pattern. "O" indicates the data point is observed; "M" indicates the data point is missing. Variables and cases from an arbitrary pattern may be rearranged to produce a monotone missingness in some cases although not for this pattern.

Multiple imputation provides one method for recovering missing data and providing accurate parameter estimates and standard errors under a MAR assumption. Another technique considered to be state-of-the-art is maximum likelihood (ML) estimation (Schafer & Graham, 2002). The EM algorithm described previously is one method to produce maximum likelihood estimates of mean vector and covariance matrix elements. In a similar way, information is "borrowed" from conditional relationships between variables in the dataset to propose final parameter estimates. A description of

ML is warranted here as this method has shown promise in providing unbiased parameter estimates and standard errors.

Full Information Maximum Likelihood (FIML) Estimation

FIML estimation is also referred to as raw maximum likelihood estimation given that a likelihood function is obtained for each individual (Graham et al., 1996). That is, all available data for an individual is incorporated in estimating parameters. Assuming multivariate normality, an individual's log-likelihood function in the presence of missing data is (Enders, 2010, p. 88):

$$\log L_i = -\frac{k_i}{2}\log 2\pi - \frac{1}{2}\log|\mathbf{\Sigma}_i| - \frac{1}{2}(\mathbf{Y}_i - \boldsymbol{\mu}_i)^T\mathbf{\Sigma}_i^{-1}(\mathbf{Y}_i - \boldsymbol{\mu}_i) \qquad (6)$$

where k_i is the number of complete data points (i.e., items or scales) for an individual and $\mathbf{Y}_i$ is that individual's vector of scores. Finally, $\boldsymbol{\mu}_i$ and $\mathbf{\Sigma}_i$ are the population mean vector and covariance matrix, respectively, *of which that individual has complete data*. That is, different values of $\boldsymbol{\mu}_i$ and $\mathbf{\Sigma}_i$ are auditioned for each missing data pattern. Sample log-likelihoods are the sum of individual log-likelihoods and are computed for *complete* and *incomplete* data patterns. The sum of all complete and incomplete sample $\log L_i$ values are used to determine the best estimates of $\boldsymbol{\mu}$ and $\mathbf{\Sigma}$ (Baraldi & Enders, 2010). In other words, the final estimates of $\boldsymbol{\mu}$ and $\mathbf{\Sigma}$ are those most likely to have produced the observed data. Under MCAR and MAR assumptions, FIML provides unbiased estimates, fewer convergence issues, and appropriate Type I error rates when compared to traditional deletion techniques (Enders & Bandalos, 2001). Like MI, statistical power may also be improved by the use of auxiliary variables when estimating FIML parameter estimates.

Several similarities between MI and FIML have been discussed. When researchers collect data via a PMD where the main missing data mechanism is MCAR for most missing data, both MI and ML procedures will produce unbiased parameter estimates and standard errors. MI and ML provide assessment practitioners with the ability to recover data missing by design *and* data that were not planned missing assuming MAR. With MI and ML, accurate parameter estimates and standard errors are not sacrificed while attempting to collect more data due to missing data. Moreover, if data quality *increases* during a PMD as a result of reduced examinee burden, these techniques serve as a gateway to the best of both worlds: a reduction in the effect of low examinee motivation on test scores without the harmful effects of traditional methods for handling missing data. It remains to be seen whether data collected via a PMD are of better quality than data collected using a full-form design, and this gap serves as an objective for this dissertation study.

Summary and Current Research Questions

Shortening the length of educational assessment and accountability tests via a planned missingness design has been proposed to reduce examinee burden while ensuring sufficient data collection for the institution. According to test-taking motivation theory, examinees consider several factors when determining the amount of effort to expend on an exam. One factor is the motivational cost of completing an exam. Cost may be high when the test is difficult and other demands for student time are high. With the use of PMDs, the arduousness of an exam may be lessened which may help sustain test-taking effort. Through the use of multiple imputation to "recover" lost information, full-form and PMD methods on the same test may be compared in several ways.

The current study has contributed to higher education assessment practice by attempting to address the issue of low examinee motivation by reducing examinee burden with a PMD. That is, students experiencing the same exam as a Long Form or as a *Short Form* via a PMD may exhibit differential examinee motivation and therefore test performance. The research questions stated here will guide this study.

1. Do aggregate total test means differ between Long and Short Forms?
2. Do aggregate total test score reliabilities differ between Long and Short Forms?
3. Are item level difficulty and discrimination indices different for Long and Short Forms?
4. Do self-reported examinee motivation (i.e., test-taking effort and perceived test importance) differ between Long and Short Forms?
5. Finally, does the relationship between test-taking effort and test performance differ between Long and Short Forms?

CHAPTER THREE

Methods

The purpose of the current study was two-fold: (a) to explore the efficacy of a planned missingness design (PMD) on sustaining or improving the psychometric quality of student learning estimates and (b) to determine if reduction of examinee burden results in higher test-taking motivation and improved performance. Specifically, a PMD was implemented on a single test to reduce examinee burden. This same test was also administered as a full-form, as it is typically administered, to collect data for assessment and accountability reporting purposes. The main hypotheses concern whether a reduction in examinee burden resulted in differences in data quality at the test-level, item-level, and examinee motivation during the test.

Participants and Procedure

Data were collected in the spring semester of 2015 from college students who had achieved sophomore or junior class status. The sample of students in the current study (N = 1,243) participated in a university-wide assessment day at a mid-sized public university in the Southeast United States. These two-hour testing sessions are proctored by at least two trained volunteers who read test instructions, pass out materials, and encourage students to give their best effort. The primary purpose of these assessment days is to measure student learning as a result of foundational General Education coursework. Results from these data are used to inform programmatic and policy changes at the institution as well as to report to state-level accountability agencies. Thus, the assessment testing context is high-stakes for university programs and personnel but low-stakes for students as their individual performance is aggregated with their peers.

Students at this institution are required to complete two assessment days during their career as college students: once as an incoming first-year student the week before fall classes begin and again in the spring after accumulating between 45 and 70 credit hours or sophomore or junior class status. The University Assessment office notifies students of their eligibility early in the spring semester. If students fail to attend assessment day, a hold is placed on the student's account. This hold prevents students from registering for future courses. Students are required to attend a make-up session at a later date to remove the hold. A very high percentage of students required to attend actually attend their second assessment day. The participation rate was very high in the current study; ≈ 90% of students required to attend spring 2015 assessment day actually attended, and after make-up assessments, the participation rate was close to 100%.

Students were randomly assigned to a testing session and room according to their student ID number. Three sessions were conducted throughout the day: Session A, B, and C. The scheduled assessments in each testing room consisted of a mixture of general education and non-cognitive measures; thus, several test configurations were established. Test configurations were created so that total testing time, including distribution of materials, did not exceed 125 minutes. Every testing room was proctored by at least two trained personnel to ensure consistency in testing procedures throughout the day. Only students who completed their assessments during their regularly scheduled session were included in this study.

Students in the current study completed one of seven quantitative and scientific reasoning test forms followed immediately by test-specific measures of examinee motivation. Students then completed various other cognitive and/or non-cognitive tests

followed by a general measure of examinee motivation concerning the entire testing session. All measures used in this study are described in the next section. All testing configurations for the current study were estimated to take between 120 and 125 minutes.

Measures

Quantitative and scientific reasoning. Quantitative and scientific reasoning ability was measured by the Natural-World test, version 9 (NW-9; Sundre, Thelk, & Wigtil, 2008). Seven forms of this test were administered: the NW-9 Long Form and six NW-9 Short Forms (A-F). The NW-9 Long Form consists of 66 items and students were given 60 minutes to complete this form. The six short forms of the NW-9 (A-F) were created as follows. The 66-item Long Form was divided into four sections of 16 or 17 items each. For example, items 1-17 formed one item set and items 18-33 formed another item set. These sections were then combined systematically to create six forms that contained between 32 and 34 items. This design is similar to the three-form design where examinees complete two-thirds of a test (Graham et al., 2006); however, examinees in the six-form design only complete one-half of a test. Students assigned to NW-9 Short Forms were given 30 minutes to complete. The item content of all seven forms is presented in Table 3. A more detailed item assignment delineation may be found in Appendix A.

Due to differences in test administration time, the NW-9 Long Form was administered in one testing room, whereas the NW-9 Short Forms were administered in two different testing rooms. The NW-9 Short Forms were spiraled throughout the assigned testing rooms so that each student received a different form than the student beside them. In addition, spiraling allowed random assignment of students to forms as

Table 3

Composition of NW-9 Test Forms and Total Testing Time

NW-9 Test Form	Item Set A Items 1-17	Item Set B Items 18-33	Item Set C Items 34-50	Item Set D Items 51-66	Item Set Order	Total Number of Items	Testing Time (Minutes)
NW-9 Long Form	O	O	O	O	ABCD	66	60
NW-9 Form A	O	O	M	M	AB	33	30
NW-9 Form B	M	O	O	M	BC	33	30
NW-9 Form C	M	M	O	O	CD	33	30
NW-9 Form D	O	M	M	O	DA	33	30
NW-9 Form E	O	M	O	M	AC	34	30
NW-9 Form F	M	O	M	O	DB	32	30

Note. An "O" indicates those items were observed on the form. An "M" indicates those items were missing from the form. Item numbers are the serial position of the items in the NW-9 Long Form.

well as ensuring approximately equal sample sizes per form. Recall that random assignment of test form is one of the tenants of PMDs and facilitates a MCAR mechanism (Graham et al., 2006). The first item on all NW-9 Short Forms asked the student to indicate which form they were completing. This information was used to determine the position and content of the items for scoring purposes and for comparison with the Long Form item responses. Test proctors were instructed to draw student attention to the first question and reiterate the importance of reporting the correct form ID. During testing, proctors circled the room to ensure that students completed this item correctly. All proctors reported that all students had indicated the correct form ID. The remainder of the test instructions for the NW-9 Long Form and the six NW-9 Short Forms were nearly identical except for the amount of testing time allowed for each version (i.e., 60 minutes or 30 minutes, respectively).

Test-taking motivation. Examinee motivation during testing was measured by the Student Opinion Scale (SOS; Sundre & Moore, 2002). Students respond to this 10-item measure using a 5-point Likert scale ranging from 1 "Strongly Disagree" to 5 "Strongly Agree." The SOS consists of two subscales: Test-taking Effort and Test Importance each comprised of five summed item responses (see Appendix B). Consistent support for a two-factor solution during a low-stakes testing context has been reported (Thelk, Sundre, Horst, & Finney, 2009). Internal consistency estimates for both the Effort and Importance subscales scores have been adequate in low-stakes contexts with students of the same age ($\alpha > .85$ for Effort; $\alpha > .82$ for Importance).

A test-specific version of the SOS immediately followed both the NW-9 Long and Short Form administrations. This measure asked students to report the effort expended on

and level of importance of *the test they just completed*. This measure was completed before students completed any additional tests during their session. After completing the test-specific SOS, students in the Long Form room completed Cognitive Test A, a 40-minute assessment. Students in the NW-9 Short Form Room 1 also completed Cognitive Test A followed by a 20-minute non-cognitive assessment. Students in the NW-9 Short Form Room 2 completed Cognitive Test B, also a 40-minute assessment followed by the same 20-minute non-cognitive assessment. All rooms then completed a 5-minute General SOS that asked students to report their motivation *across the entire testing session*. Table 4 describes the test configurations in each room as well as total testing time allowed.

Table 4
NW-9 Room Test Configurations

Testing Room	First Test	Second Test	Third Test	Fourth Test	Testing Time (Minutes)
Long Form Room	NW-9 Long Form	Cognitive Test A	General SOS		125
Short Form Room 1	NW-9 Short Forms	Cognitive Test A	Non-Cog Test	General SOS	120
Short Form Room 2	NW-9 Short Forms	Cognitive Test B	Non-Cog Test	General SOS	120

Note. Testing times allowed for each test are as follows: NW-9 Long Form = 60 minutes; NW-9 Short Forms = 30 minutes; Cognitive Test A = 40 minutes; Cognitive Test B = 40 minutes; Non-Cog Test = 20 minutes; General SOS = 5 minutes. These same test configurations repeated for Session A, B, and C.

Data Analysis

The research questions in the current study may be categorized into three sections: (a) test-level comparisons (RQs 1 and 2), (b) item-level comparisons (RQ 3), and (c) the examinee experience (RQs 4 and 5). All analyses addressed differences between a Long Form and PMD in these three areas. The first set of research questions (1 & 2) compared total test means and reliability estimates between the NW-9 Long and Short Form

conditions. Research question 3 explored if the data collection design resulted in differences in item-level difficulty and discrimination parameters. The final set of research questions (4 & 5) explored the differences in examinee motivation during the test. Specifically, average self-reported examinee Test-taking Effort and Test Importance scores was compared between the Long Form and Short Form conditions. Additionally, the relationship between Test-taking Effort and test performance was compared between Short and Long Form conditions given prior research concerning this effect (Cole et al., 2008; Huffman et al., 2011; Silm et al., 2013; Sundre & Kitsantas, 2004; Wolf et al., 1995). Because PMDs result in data that are MCAR, the choice between MI and ML should not affect final estimates, so MI was used to account for both planned and unplanned missingness in Short Forms and the Long Form, as well as in SOS item responses. All planned and unplanned missing data was recovered using MI. The imputation phase in the current study was somewhat different from typical imputation phases. The specific steps of the imputation phase of the MI model are now detailed.

Imputation model specification. When researchers are interested in imputing missing data in a dataset, the researcher includes any variables of interest to the current study that have missingness (Enders, 2010). These missing values are then replaced with imputed values from the imputation model. Because the current study has two separate testing conditions (Short vs. Long Form), these data were imputed separately. Data from all NW-9 Short Forms were combined and used in the imputation model for the Short Form condition. Two separate imputation models were run to prevent the influence of item relationships within the Short Form condition from influencing the imputation of values in the Long Form condition and vice-versa. Moreover, if an assessment

practitioner only collected data via a PMD, only short forms would be included in the imputation model. Item-level imputation was used given the improvement in estimation precision when compared to scale-level imputation (Gottschall et al., 2012). Item-level imputation is also necessary given that it is these item responses, imputed and observed, that were compared in research question 3.

Each individual was presented with the test-specific SOS after their assigned Short Form; however, some unplanned missing data in the SOS was expected. Both complete and incomplete NW-9 Short Form and SOS item responses were imputed in one imputation model. By including all variables of interest in the current study in the same imputation model, relationships among these items are preserved (Enders, 2010). This specification of the imputation model is necessary given that research question 5 addresses the relationship between Test-taking Effort scores and total test performance. The rate of planned missingness for the Short Form data was 50% and this missing data rate was expected to be higher given some unplanned missingness in the NW-9. Given this amount of missing data, 40 imputations were generated to ensure sufficient statistical power and accurate standard errors (Graham et al., 2007).

Observed item responses were scored prior to imputing missing responses. Both Long and Short forms were scored identically. Out-of-range responses (i.e., selecting option “D” when answers to the item only included “A” through “C”) were scored as missing. However, no out-of-range data was observed in either condition. Scoring an omitted item as incorrect is not recommended when examining uniform (Finch, 2011b; Robitzsch & Rupp, 2009) or non-uniform DIF (Finch, 2011a). All unplanned missingness

was left as missing rather than scored incorrect and these values were also imputed. Only attempted items were scored.

PMDs, particularly the design used in this study, result in an arbitrary missingness pattern. With this kind of missing data, the optimal imputation model is linear imputation with no rounding (Allison, 2005). That is, missing item responses on the NW-9 Short Form and SOS items were predicted from all complete item responses and these predicted values were not rounded. For this cognitive test, data were imputed using scored item responses. Good auxiliary variables are moderately to highly correlated with the missing values (Collins et al., 2001). Therefore, items on the same test should correlate with each other and serve as good auxiliary variables when some item responses are missing. A fully "inclusive" strategy of adding many auxiliary variables was not used given the unknown missingness mechanism of the unplanned missing data.

Although not planned by design, the Long Form data was expected to contain some missing data. Missing data in the NW-9 and SOS items when data was collected using the Long Form was imputed in a separate MI model from the Short Form data. Although the total amount of missing data was much smaller in the Long Form than the Short Form, the purpose of this study was the compare the results of these two data collection methods. Thus, a total of 40 imputed datasets were imputed for the Long Form data to be compared to the Short Form data.

Some research questions sought to model the differences in relationships between the two conditions; therefore, a dummy variable (G) was added to each imputation to determine from which group the data originated (0 = Long Form; 1 = Short Form). Then, the individually imputed datasets for the Long Form and Short Form designs were

concatenated by imputation. That is, the first imputed dataset for the Long Form data was concatenated with the first imputed dataset from the Short Form data. This process was repeated for all 80 imputed datasets. In this way, a final set of 40 imputed and matched datasets was used for analysis. At this point, total test and subscale scores were computed for each individual.

Research question 1: Total test mean scores. Do aggregate total test means differ between Long and Short Forms? Total test performance means for both test form conditions (i.e., Long and Short Form) were calculated across the datasets. MI standard errors, confidence intervals, and diagnostic indices (adjusted degrees of freedom and FMI) were calculated from the within- and between-imputation variance components. An independent-samples t-test was performed on each of the 40 multiply imputed datasets comparing mean total test scores between the two testing conditions. The results of these tests (i.e., the analysis phase) were combined using D. B. Rubin's (1987) rules described in Chapter 2 to obtain a single test of mean differences. A Levene's test was conducted with each independent-samples t-test to assess the tenability of the homogeneity of variances assumption.

Research question 2: Total test score reliability. Total test score reliability was compared between the Long Form and Short Form conditions. Coefficient α was computed within each multiply-imputed dataset by condition using the standard formula (Equation 10).

$$\alpha = \frac{k}{k-1}\left(1 - \frac{\sum \sigma_i^2}{\sigma_x^2}\right) \tag{10}$$

The 40 coefficient α values were then averaged by condition. No transformation to the α coefficients was conducted prior to averaging for concern of adding noise to the comparison (Enders, personal communication, September 23, 2014). The Cronbach's α for the Long Form data was computed for each multiply-imputed dataset using the same formula and averaged across datasets. These two coefficients were tested for significance using an extension of Feldt's (1969) formula for independent samples (Feldt, Woodruff, & Salih, 1987).

Research question 3: Item-level differences in difficulty and discrimination. Item-level differences in difficulty and discrimination between the Long Form and the Short Forms may further describe the effects of a planned missingness design on a single test. Logistic regression DIF procedures were used to test differences in item parameters between the two form conditions after controlling for ability (Swaminathan & Rogers, 1990). That is, the probability of correctly answering each item was modeled by individual ability (X_i), grouping variable (G), and the interaction between ability and grouping variable (X_iG). This model is presented in Equation 11. In this application of DIF, the grouping is form condition (Long Form vs. Short Form).

$$P(u=1) = \frac{e^{[\beta_0+\beta_1 X_i+\beta_2 G+\beta_3(X_iG)]}}{1+e^{[\beta_0+\beta_1 X_i+\beta_2 G+\beta_3(X_iG)]}} \quad (11)$$

Including total test score (X_i) control for ability when comparing how group membership relates to item score. Conceptually, DIF is present when the predicted probability of the same item for individuals with the same ability is different across two groups. That is, after controlling for ability, significant differences in item response probability explained by the grouping variable will be evidenced by a significant β_2

value. A nonzero β_2 and a zero β_3 indicate uniform DIF in that the item appears more difficult for one group than the other, with the difference being the same at all levels of ability. A nonzero β_3 value indicates the presence of non-uniform DIF, or an inference that the item discrimination parameters differ between the groups. With non-uniform DIF, there is a difference between groups in item response, but the difference changes in magnitude or direction across the ability continuum. As with research question 1, model parameters were combined using D. B. Rubin's (1987) rules. Therefore, final results for each item represent the average logistic regression model across the imputed datasets.

Research question 4: Average test-taking motivation. Differences in the psychometric properties of items and forms in a planned missingness design were hypothesized to be explained by differences in an examinee's experience. In particular, examinee motivation, as measured by the Test-taking Effort and Test Importance subscales of the SOS, may explain differences in the examinee experience when participating in a PMD. Mean Test-taking Effort and Test Importance were compared between the two test conditions. A Levene's test of the homogeneity of variances was run for each test and compared to assess the tenability of this assumption.

Research question 5: Relationship between test-taking effort and test performance. The relationship between Test-taking Effort and test performance has been well established. However, this relationship has not been compared between the Long Form of an instrument and a Short Form condition via a PMD. A multiple regression model was run to test the difference in the relationship between scores on the Effort subscale and NW-9 total scores (Equation 12).

$$\mathrm{Y} = \beta_0 + \beta_1 \mathrm{Effort} + \beta_2 G + \beta_3 (\mathrm{Effort} \times G) + e \quad (12)$$

The dependent variable, NW-9 total test scores, was regressed on Test-taking Effort subscale scores (β_1), test condition or group (β_2), and the interaction between Test-taking Effort scores and test condition (β_3). A nonzero β_3 would indicate that the relationship between Test-taking Effort scores and total test scores differs between the two test condition groups.

CHAPTER FOUR

Results

This chapter will provide findings for all study results and is organized to describe the samples in each test form condition, present the imputation models, and then detail the findings for each research question. All data were read into SAS 9.3 and PROC MI was used to run the multiple imputation models. Analyses varied by research question; however, most results were combined using PROC MIANALYZE which automates the combinations of parameter estimates and standard errors following Rubin's (1987) rules. Recall that six research questions were proposed. Research questions 1 and 2 compare total test score means and reliability estimates between the Short and Long Forms. Research question 3 examines differences in item parameter estimates. Finally, research questions 4 and 5 introduce the examinee motivation variables to ascertain the influence of a planned missingness design (PMD) on examinee motivation.

A total of 216 students were randomly assigned to the NW-9 Long Form and 1,030 students were randomly assigned to one of six NW-9 Short Form conditions. NW-9 Short Forms were spiraled so that each student was administered a randomly chosen form. This design resulted in about equal sample sizes across the forms (see Table 5). Three students in the Short Form condition completed their student ID and name on their answer sheet but did not answer a single item. Without any data for these three students, they were removed from further analyses. In addition, some students did not respond to all items administered. That is, some students did not respond to all items on the NW-9 Long Form or to all administered items on the NW-9 Short Form. Unplanned missingness was very small across the Short Forms with students not responding to less than 1

Table 5
Sample Size and Missingness Rates by NW-9 Test Form

	N	Number of NW-9 Items Administered	NW-9 Planned Missingness	NW-9 Unplanned Missingness	SOS Unplanned Missingness
Long Form	216	66	0%	< 0.01%	2.36%
Short Form (All)	1,027	33	50%	0.26%	2.32%
NW-9 Form A	173	33	50%	0.68%	2.20%
NW-9 Form B	172	33	50%	0.14%	1.16%
NW-9 Form C	168	33	50%	0.41%	1.73%
NW-9 Form D	171	33	50%	0.87%	4.44%
NW-9 Form E	170	34	≈ 52%	0.61%	2.47%
NW-9 Form F	173	32	≈ 48%	0.36%	1.91%

Note. Unplanned missingness is the average number of omitted or not reached items to total administered items. All forms and conditions were administered 10 SOS items. The Short Form (All) condition describes all NW-9 Forms A-F when combined.

Table 6
Demographic Information by Test Form Condition

					Ethnicity Percentages						
	N	% Female	Mean Age	Mean GPA	White	Black	Asian	Hispanic	American Indian	Pacific Islander	Not Specified
Long Form	216	56.13	20.21	2.82	87.26	5.66	4.72	5.66	1.89	--	1.89
Short Form	1,027	63.05	20.15	2.93	84.96	4.68	8.86	6.67	0.60	0.60	2.49

Note. Students are allowed to indicate more than one ethnicity. Thus, ethnicity percentages may exceed 100%.

administered item, on average. Unplanned missingness in Table 5 was calculated by dividing the average number of omitted or not reached items by the total administered items. For example, students who were administered the NW-9 Short Form A did not respond to 0.225 NW-9 items on average. Form A included 33 administered NW-9 items for an average missingness rate across students of 0.68%. Mean rates of unplanned missingness for both the NW-9 and SOS scales are found in Table 5. Notably, NW-9 missingness was more common in the Short Form condition compared to the Long Form condition and missingness was more pronounced in the SOS compared to the NW-9 in both conditions. However, these rates of unplanned missingness are very small.

At this point and as described in Chapter 3, all NW-9 Short Forms were concatenated to result in a final sample size of 1,027. Demographic information for both the NW-9 Long and Short Form conditions may be found in Table 6. Notably, random assignment to condition appears to have resulted in a similar demographic profile in both conditions. Tests were scored using the same test key. Complete item responses on both forms were scored "0" for incorrect and "1" for correct. Surprisingly, there were no out-of-range responses in either form condition. Importantly, omitted or not reached items were left missing in the scored item response vector rather than scored incorrect. Thus, omitted or not reached items were imputed along with the planned missing responses.

Multiple Imputation Model Convergence

Short form. The Short Form multiple imputation model was specified using all available data from the six short forms. That is, all scored item responses to the NW-9 Short Forms (A-F) and SOS were concatenated into a single dataset including planned and unplanned missingness. Starting values for the EM algorithm (i.e., item means and

covariances) were provided using all available data. The EM algorithm converged in 1,724 iterations using the default convergence criterion for SAS 9.3 of .0001. That is, between the 1,723rd and 1,724th iteration, the absolute value change for each parameter estimate was less than .0001. Given the large amount of missing data, the number of iterations was not surprising. When the EM algorithm does not converge within 200 iterations, Graham (2012) suggests that increasing the number of iterations for the EM algorithm is preferred instead of increasing the convergence criterion.

The EM estimates were used as starting values for the MCMC chains. The multiple chains option in PROC MI was used. That is, each imputed dataset was a result of 40 separate MCMC chains. All observed, scored item responses to both the NW-9 and SOS were used as auxiliary variables for missing item responses. A non-informative prior was first used; however, the solution did not stabilize after 265 iterations. Recall from Chapter 2 that the P-step parameter estimates in the MCMC procedure must converge (i.e., stabilize) in a distributional sense. SAS reported that the initial covariance matrix was singular (i.e., linearly dependent variables). At this point, SAS allows for the user to specify prior information (i.e., data from a similar cohort with less missing data) or a ridge prior. To avoid possible biases from using prior information from another sample, a ridge prior was used to assist in convergence.

Conceptually, a ridge prior adds imaginary cases (equal to the hyperparameter *d**) to the dataset that have complete data but with 0 covariance between variables. Consequently, relationships between variables are very slightly biased. With no cases in the Short Form condition with complete data, this procedure should assist in

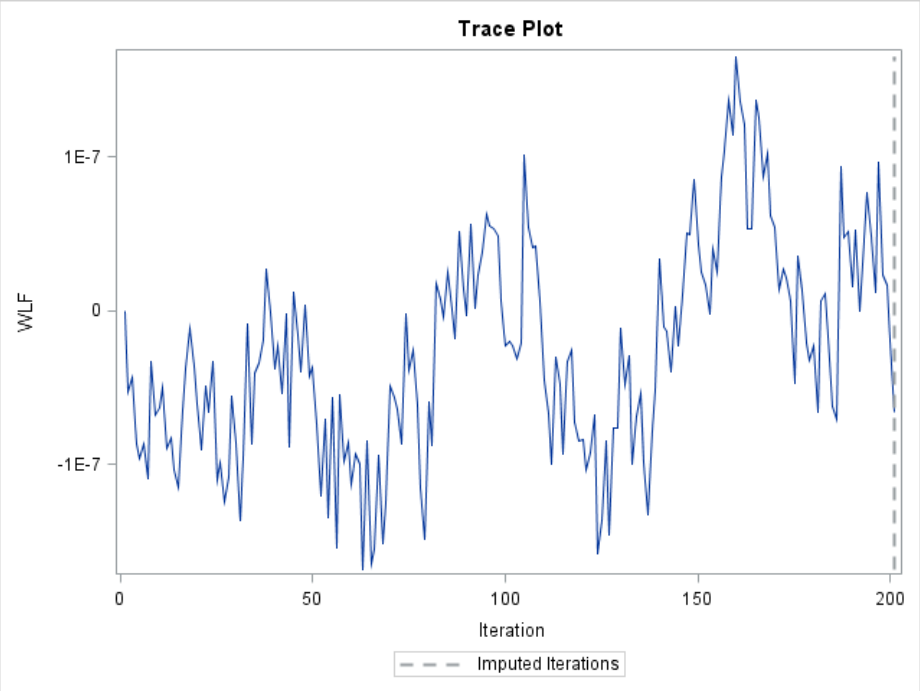

Figure 3. The worst linear function (WLF) time-series plot for the Short Form imputation model. The plot indicates stable convergence within the 200 iteration burn-in period. Note the very small numbers on the y-axis.

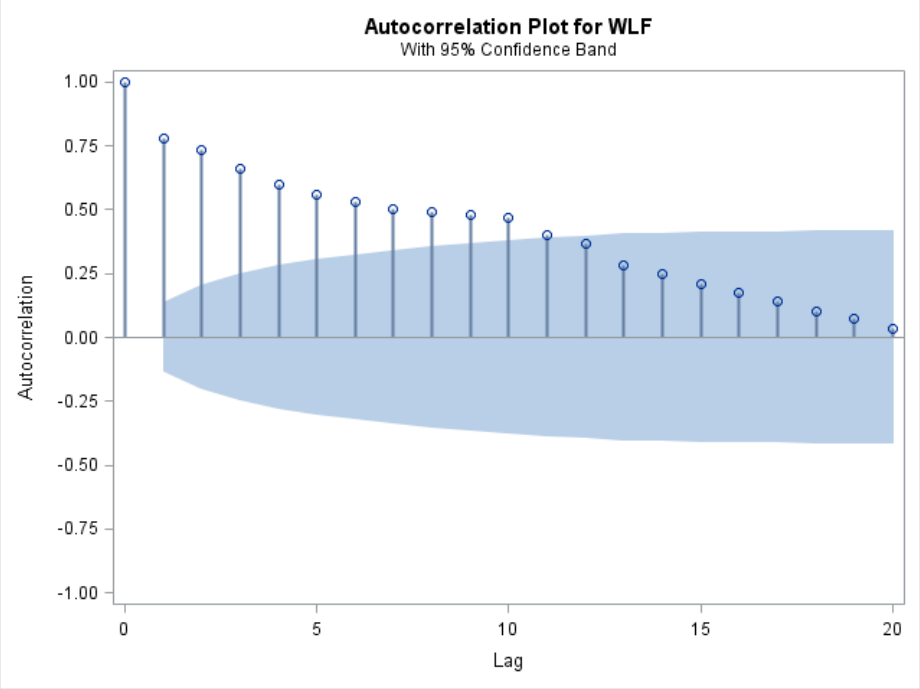

Figure 4. The worst linear function (WLF) autocorrelation plot for the Short Form imputation model. The plot indicates that imputations become independent somewhat quickly with autocorrelations falling within sampling error of 0 within about 12 imputations.

convergence. Methodologists have suggested specifying a d^* value that equates to no more than 1% of the total sample size (Graham, 2012). With a ridge prior of $d^* = 10$, the MCMC procedure converged in 646 iterations. The worst linear function (WLF) time-series and autocorrelation function plots for the Short Form imputation model are found in Figures 3 and 4, respectively. The WLF time-series plot indicated no dependencies (i.e., trends veering off) among the datasets within the 200 burn-in iterations. The WLF autocorrelation plot suggests that imputed datasets appear to be independent (i.e., within sampling error of 0) within 12 imputations. A safe margin of 200 burn-in iterations were specified for each MCMC chain before an imputation was drawn. A total of 40 datasets were generated from these parallel MCMC chains. All Short Form datasets were given a value of "1" for the grouping variable at this time.

Long form. Although there were no missing data by design, it will be recalled from Table 5 that the Long Form condition resulted in minuscule amounts of unplanned missing data between both the NW-9 and SOS items. The imputation model for the Long Form data incorporated all observed and missing NW-9 and SOS items. Items were scored identically to the Short Form condition with out-of-range and omitted items scored missing rather than incorrect. Starting values for the EM algorithm (i.e., item means and covariances) were provided using all available data in the Long Form condition. The EM algorithm converged in 22 iterations, as expected much faster than the Short Form condition.

The EM estimates were used as starting values for the parallel MCMC chains. As in the Short Form condition, multiple chains were used and all observed item responses were used as auxiliary variables for missing item responses. A non-informative prior was

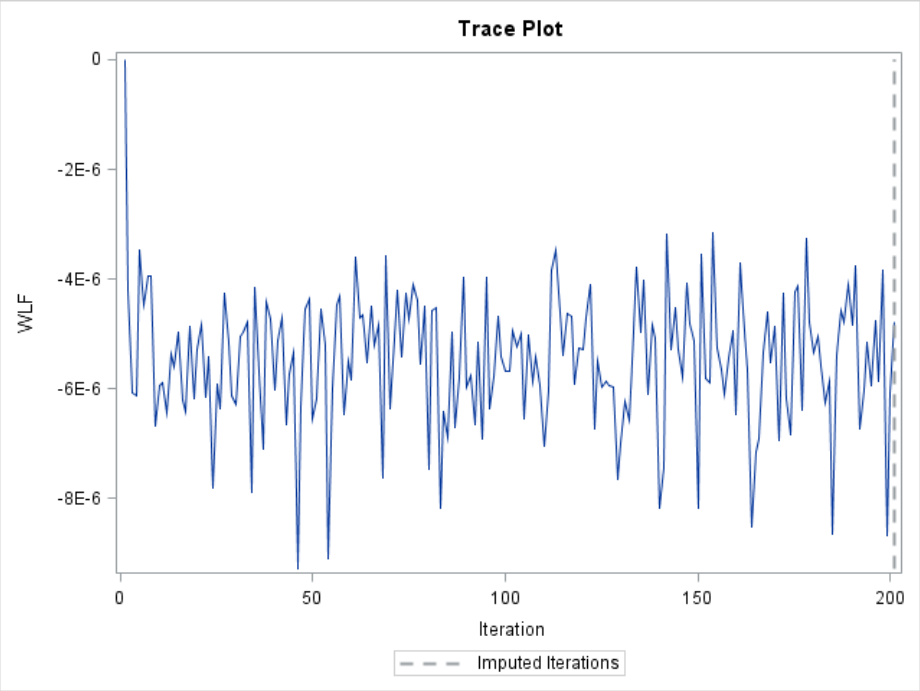

Figure 5. The worst linear function (WLF) time-series plot for the Long Form imputation model. The plot indicates stable convergence within the 200 iteration burn-in period.

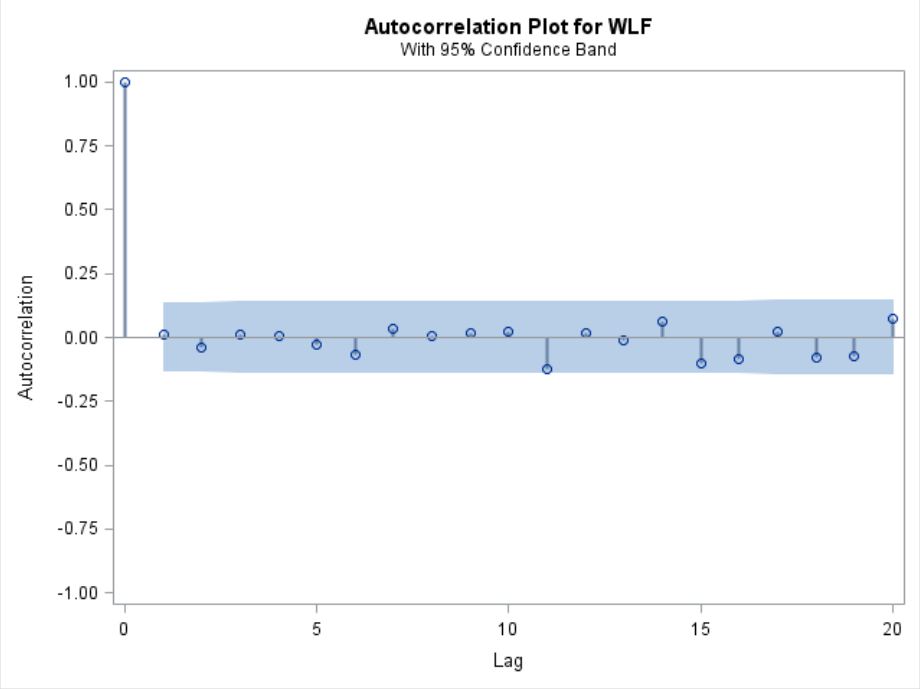

Figure 6. The worst linear function (WLF) autocorrelation plot for the Long Form imputation model. The plot indicates that imputations become independent very quickly with autocorrelations falling within sampling error of 0 within about 2 imputations.

used and the solution stabilized in 9 iterations. The worst linear function (WLF) time-series and autocorrelation plots for the Long Form imputation model are located in Figures 5 and 6, respectively. Like the Short Form imputation model, both plots suggest no convergence issues. Independence between datasets was attained much earlier in the Long Form with autocorrelations falling within sampling error of 0 within 2 iterations. A total of 200 burn-in iterations were specified and 40 datasets were generated. All Long Form datasets were given a value of "0" for the grouping variable at this time.

All research questions compare parameter estimates between Short and Long Form conditions. Thus, these multiply-imputed datasets needed to be combined to address the research questions. All 40 Short and Long Form *imputed* datasets were concatenated at this time. Each dataset contained the imputation number ($m = 1, 2 \ldots 40$) as well as a grouping variable specified earlier. The model used to answer each research question was estimated within each imputation. For example, to calculate total test score means, the mean from the Short Form imputation 1 was compared to the mean form the Long Form imputation 1. Each model was run on each imputed and concatenated dataset (i.e., analysis phase) and their results combined (i.e., pooling phase) to provide a single, final analysis. The final results for each research question are now detailed.

Research Question 1: Do aggregate total test means differ between Long and Short Forms?

Mean group total scores on the NW-9 are presented in Table 7. Due to MI, these group-level total test scores are on the same metric of number of correct items. However, with no rounding, scored item scores may exceed 1 and may lead to a sum total score greater than 66. Mean total test scores were statistically significantly higher in the Short

Table 7

Total Test Performance Means and Reliability by Test Form Condition

								95% CI				
	N	Mean (SD)	SE	v_1	FMI	Skew	Kurtosis	LB	UB	α	*t*	*d*
Long Form	216	47.64 (8.65)	0.551	1238.8	< 0.001	-0.666	0.219	46.557	48.718	.860	2.26*	.172
Short Form	1,027	49.02 (7.97)	0.615	1177.1	0.028	-0.798	0.950	47.818	50.230	.827		

Note. Point estimates in this table (SD, Skew, and Kurtosis) are the mean value across imputations. Cohen's *d* represents a standardized mean difference and is interpreted as the difference in standard deviation units. SD = standard deviation; SE = standard error; v_1 = adjusted degrees of freedom; FMI = fraction of missing information; LB = lower bound; UB = upper bound; α = coefficient alpha; *d* = Cohen's *d*.

* $p < .05$.

Form condition than the Long Form condition (t = 2.26, p = .024). This raw mean difference of 1.39 points resulted in a standardized effect size of .172. According to Cohen's (1980) guidelines, this effect size was practically small (< .20). Test score variance was smaller in the Short Form condition than the Long Form condition. That is, on average, students in the Short Form condition performed more similarly to each other than in the Long Form condition. Levene's test of homogeneity of variances were conducted between the forms. Recall that the results for RQ1 were to pooled results of 40 separate analyses. When examining the Levene's test results, variances were equal in 35 of the 40 tests.

Research Question 2: Do aggregate total test score reliabilities differ between Long and Short Forms?

Coefficient α was used to assess the internal consistency of the imputed datasets. Forty separate α values were calculated for each dataset and the average of these estimates within each form condition is reported in Table 7. Total test reliability was somewhat lower in the Short Form condition than the Long Form condition: .83 and .86. This small difference might be expected given the reduced variability of Short Form test scores. When performances of a sample of individuals are more homogenous (i.e., smaller variance), estimates of reliability generally decrease. A freeware program for testing coefficient α values was used to evaluate the statistical significance of this difference (Lautenschlager & Meade, 2008). Using the formula for independent samples (Feldt, Woodruff, & Salih, 1987), the two coefficients were not statistically significantly different from each other (χ^2 (1) = 3.68, p = .055). Although total test scores were different, test reliability coefficients were similar between the two form conditions.

However, item-level differences may remain that are not uncovered when looking at total scores.

Research Question 3: Are item level difficulty and discrimination indices different for Long and Short Forms?

Item-level differences between the forms were assessed using a differential item functioning (DIF) framework. In typical DIF analyses, item parameters are compared between two or more known groups (i.e., gender, ethnicity). If two individuals from different groups but with the same ability have different probabilities of responding correctly to an item, this difference would constitute noteworthy DIF. In this application, DIF analyses are used to examine differences in item parameters between randomly assigned test form conditions after controlling for ability. Therefore, differences in item parameters are assumed to be the result of the two test form designs if two individuals with the same ability perform differently on an item. Classical test theory (CTT) item difficulty values (p-values) are depicted graphically in Figures C1 and C2. However, it should be noted that these differences do not control for ability, so a fair comparison necessitates a procedure that accounts for total ability.

Logistic regression (LR) analyses were used to assess differences in item difficulty and item discrimination between the Short and Long Form conditions (Swaminathan & Rogers, 1990). LR is easily implemented in SAS and results from multiple datasets may be combined using PROC MIANALYZE. First, the total scores and the grouping variable were grand-mean centered. This was done to prevent confounding of the main effect and the interaction. To test item difficulty differences, the model in Equation 11 with only total test score (X_i) and group membership (G) was run

for each item. Significant differences in CTT item difficulty were noted in nine of the 66 items (13.6%) on the NW-9 (items 2, 3, 18, 20, 26, 30, 35, 38 and 57). Recall that the grouping variable (originally coded 0 and 1) was grand mean-centered but remained one unit apart. This allowed for the main effect of form to be converted to the odds ratio (DeMars, 2009). Odds ratios are necessary to produce LR DIF effect sizes. To provide an effect size, these odds ratios were converted to the delta metric ($\hat{\Delta}_{LR}$) and were categorized using the ETS classification system (Monahan, McHorney, Stump, & Perkins, 2007). There are three DIF categories for the delta metric. Items are placed in category "A" when 1 > $|\hat{\Delta}_{LR}|$, category "B" when $|\hat{\Delta}_{LR}|$ > 1 but < 1.5, or category "C" when $|\hat{\Delta}_{LR}|$ > 1.5. Category A and B items are considered negligible DIF or slight to moderate DIF, respectively. Category C items are considered moderate to large DIF items and should be investigated further for item bias potential. All nine items mentioned previously were in category "C" which indicated significant DIF. Two items (items 2 and 18) were easier for the Short Form group whereas the rest of the items (3, 20, 26, 30, 35, 38 and 57) were easier for the Long Form group. Importantly, there appeared to be no systematic differences between the test forms on item difficulty, particularly in relation to item position.

By including total test score (X_i), group membership (G), and their interaction (X_iG) in the model, differences in item discrimination parameters between the form conditions were explored. The results of these analyses suggest that item discrimination differences are less common than observed for item difficulty. Only four items (6%) on the 66-item NW-9 resulted in significant DIF (items 14, 37, 38, and 57). Items 37, 38, and 57 were more discriminating in the Short Form group than the Long Form group.

Like item difficulty differences, there was no discernable systematic pattern of non-uniform DIF.

Both uniform and non-uniform DIF did not appear to be a function of item position. That is, if examinees in the Long Form condition were fatigued over the entirety of the test, item parameters would be systematically different in later test items (e.g., beyond item 50). However, this pattern was not observed. In fact, no systematic pattern was observed in either uniform or non-uniform DIF between the two forms. In summary, recall that a potential contribution of this dissertation was to consider whether planned missing data designs might have an impact on the psychometric quality of resulting scores. This was investigated at both the total test score and item levels. The analyses thus far suggest that the purported gains in time and administrative efficiency did not negatively impact test results. In fact, total test score means were significantly increased, with a fairly small effect size gain. Further, reliability was not greatly reduced. While item level differential functioning was observed for 9 item means, the pattern of difficulty was split across the Long and Short forms with no systematic item placement effects observed. Item level discrimination indices resulted in only four significant items, three of which resulted in greater discrimination on the Short Form with no systematic item placement effects. The second stage of analyses examined whether PMDs might result in enhanced examinee motivation, which has been shown to impact test performance.

Research Question 4: Do self-reported examinee motivation variables differ between Long and Short Forms?

Recall that both NW-9 and SOS items were included in the same imputation model for each form condition. Therefore, item-level imputation was performed on

Table 8

Test-taking Effort and Importance Means and Reliability by Test Form Condition

					95% CI				
	Mean (SD)	SE	v_1	FMI	LB	UB	α	*t*	*d*
Test-taking Effort									
Long Form	18.64 (3.78)	0.245	1192.4	.029	18.161	19.123	.842	1.53	.115
Short Form	19.05 (3.51)	0.269	1205.9	.018	18.525	19.580	.841		
Test Importance									
Long Form	15.23 (4.00)	0.281	1185.4	.025	14.676	15.777	.803	1.53	.116
Short Form	15.70 (4.08)	0.308	1193.3	.023	15.094	16.303	.840		

Note. Test-taking Effort and Test Importance subscale scores can range from 5 to 25 with higher scores indicating greater expended effort and perceived importance. Cohen's *d* represents a standardized mean difference and is interpreted as the difference in standard deviation units. SD = standard deviation; SE = standard error; v_1= adjusted degrees of freedom; FMI = fraction of missing information; LB = lower bound; UB = upper bound; α = coefficient alpha; *d* = Cohen's *d*.

missing SOS items within each condition. After imputation, subscale scores were computed for each individual within each multiply-imputed dataset. Subscale means between the test form conditions on Test-taking Effort and Test Importance scores are presented in Table 8. Mean reported Effort was higher in the Short Form condition, although this difference was not statistically significant ($t = 1.47$, $p = .135$). A standardized effect size of .115 was computed for the Effort means. Perceived Test Importance followed a nearly identical pattern between the groups. Although reported Test Importance was higher for the Short Form condition, this difference was not statistically significant ($t = 1.57$, $p = .116$). A standardized effect size of .116 was computed for the Importance means. Both effect sizes for the motivation variables would be considered practically small (< .20) according to Cohen's (1980) guidelines. The homogeneity of variances assumption was established as all Levene's tests were nonsignificant between the forms for both Effort and Importance. Internal consistency estimates for both subscales were also computed (see Table 8). Between the forms, coefficient α estimates were not statistically significantly different for the Test-taking Effort (χ (1) = 0.008, $p = .927$) or Test Importance (χ (1) = 2.62, $p = .105$) scales.

Research Question 5: Does the relationship between test-taking effort and test-performance differ between Long and Short Forms?

Although no mean difference between the form conditions on Test-taking Effort was found, variance shared between test performance and Test-taking Effort may still differ between the groups. The influence of Test-taking Effort on test scores (and not Test Importance) was primarily examined given its theoretical and established relationship (Abdelfattah, 2010; Cole et al., 2008; Waskiewicz, 2011). Moreover, a reduction in the

variance shared between effort and test performance could be indicative of a reduction in construct-irrelevant variance that may be present in these test scores.

When examining the bivariate relationship between Test-taking Effort and test performance, the relationship is somewhat weaker in the Short Form group (r = .399) than the Long Form group (r = .424). A multiple regression equation was conducted to test the magnitude of this relationship between the groups (research question 5). The dependent variable was total test scores on the NW-9 and the independent variables were Test-taking Effort scores, form condition, and the interaction between Test-taking Effort and condition. All variables were standardized prior to running the model. The results of this regression equation are found in Table 9.

Table 9
Multiple Regression Predicting NW-9 Scores from Test-taking Effort, Test Form Condition, and their Interaction

					95% CI			
	B	SE (*B*)	v_1	FMI	LB	UB	*t*	*p*
Intercept	< 0.00	0.028	697.7	.140	-.054	.055	0.02	.98
Effort	0.40	0.028	690.8	.142	.348	.458	14.38	< .01
Group	0.05	0.027	1140.8	.039	-.006	.098	1.75	.08
Effort x Group	-0.01	0.026	1037.5	.064	-.061	.039	-0.42	.67

Note. Mean model R^2 = .164. Group was coded "0" for the Long Form condition and "1" for the Short Form condition. *B* = standardized regression coefficient; SE (*B*) = standard error of regression coefficient; v_1 = adjusted degrees of freedom; FMI = fraction of missing information; LB = lower bound; UB = upper bound.

Both the multiplicative interaction (p = .67) and the main effect of group (p = .08) were nonsignificant. The main effect of Effort scores was significant (p < .01). As many studies have previously reported, Test-taking Effort scores were positively related to test performance, but this relationship did not significantly differ between the groups taking the Long or Short forms of the test. This result suggests that test scores obtained via a

PMD design can be predicted with similar efficiency as when the test is given as a Long Form. The effect was in the correct direction as the PMD did reduce the variance shared between these two variables slightly (2% less). Coupled with the results of the previous research questions, very few effects in data quality were found between the form conditions except for an increase in overall test performance. No findings suggested a degradation in data quality from using a PMD in a higher education assessment context. Any small effects that were observed favored the PMD over the Long Form design.

Summary

The results of this study highlight the differences in test and item-level scores, reliability, and examinee motivation between a PMD (Short Form) and a full-form design (Long Form). Overall, very few differences were found from the current analyses. Examinees in the Short Form condition performed better than examinees in the Long Form condition. Although the effect size was small, this difference is very meaningful for practitioners who desire to show accurate student learning. Examinees in the Short Form condition also reported similar Test-taking Effort and Test Importance to the Long Form condition, and their test scores shared 2% less variance with self-reported Effort. All effects were practically small, but these small differences were in the desired direction for the Short Form. Moreover, a *reduction* in data quality was not observed at the test or item level. The PMD used in this study resulted in data quality very similar to that of a full-form test while reducing testing time by one-half. There are several implications of these findings for assessment practitioners concerning measurement quality and assessment policy that will be discussed in Chapter 5.

CHAPTER FIVE

Discussion

This study compared data quality of two test form designs used to collect higher education assessment and accountability data. Specifically, a test of scientific reasoning was administered as both a Long Form and as a Short Form via a PMD that incorporated 50% missing data. MI procedures were used to impute and analyze all data. Although PMDs are most often used to increase data collection efficiency by reducing testing time, the current study explored whether data quality increased or remained stable when compared to a full-form design. Higher education assessment practitioners typically examine test descriptive statistics and reliability coefficients to report student learning and data quality. Therefore, these parameter estimates and coefficients were the focus of the first test-level study stage. Item parameters have been shown to differ with position change, and this research served as an impetus for the second, item-level investigation stage. Finally, examinee motivation variables were compared between the two form conditions. The low-stakes testing context of higher education assessment introduces validity concerns because of reduced examinee motivation, particularly test-taking effort. A reduction in examinee burden was proposed as a possible technique to increase examinee motivation and counter its negative influence on test performance. This discussion will be framed by these three stages of study followed by study limitations, implications for higher education assessment, and recommendations for future research.

Test-Level

Effect on group-level estimates. Group mean test performances were first compared between the Short and Long Forms of the scientific reasoning test. The group

mean in the Short Form condition was significantly higher than in the Long Form condition. Although this difference was statistically significant, the effect size was practically small (d = .172). Higher education practitioners desire to accurately capture student knowledge and there is some doubt surrounding our current estimates. If long and arduous tests are resulting in underestimates of true ability, it is important to identify how and why this is occurring. Low examinee motivation has been frequently cited as a predictor of test performance in low-stakes contexts. The PMD in this study may have provided a slight correction to this difference between two randomly-assigned samples. Though the effect size is considered 'small', these effect sizes must be interpreted within context. In assessment situations that are considered high-stakes for institutions and educational programs, differences in group mean performance estimates are of great importance. Students were randomly assigned to Long or Short Form conditions resulting in samples similar in scientific reasoning ability, prior coursework, age, and other variables. Using this quasi-experimental design, there are few variables other than sampling that could contribute to this difference. Therefore, the difference in group mean performance was not a function of ability but rather due to the data collection design. The current study results suggest that PMDs can contribute not only to enhanced data collection efficiency but may lead to increased student learning outcomes.

Even though an important effect in group mean performance was observed, the PMD design would still be an attractive option if group mean performance between the form conditions was equal. The PMD coupled with a modern missing data handling technique did not reduce the measurement precision of this group-level information. In fact, MI resulted in scores from the PMD that were on the same metric as the full-form

test allowing practitioners to compare these estimates with previous administrations of the same instrument. This means that practitioners may change to a PMD with confidence that the parameter estimates will be comparable to those previously obtained. More importantly, the results of this study indicate a likelihood of higher group estimates possibly due to a reduction in examinee burden and fatigue.

PMDs have been implemented in K-12 contexts such as during the NAEP exams (Zwick, 1991). However, these designs have not been used in higher education assessment, and no study was found linking PMDs as a potential solution to the examinee motivation problem. With a paucity of studies using PMDs in higher education assessment, the group mean results are difficult to compare to prior research. This lack of an established literature should stimulate researchers to formulate new studies exploring the utility of these designs. This study serves as an initial exploration of these designs' efficiency and utility.

Reliability of test scores. The internal consistency estimate of the Short Form condition ($\alpha = .827$) was lower than the Long Form condition ($\alpha = .860$). Although the findings for RQ 2 suggested that these two coefficients were not significantly different from each other, some readers may be concerned with a reduction in coefficient α when considering PMDs. There are a few possible explanations for this difference. Coefficient α has been shown to be spuriously deflated on scored right-wrong tests when the test is *speeded* (Attali, 2005). During speeded tests, test-takers do not have enough time to complete all items. Without sufficient time, students may omit responses or rapidly guess on the remaining items. These guesses appear as random "noise" in the data and will attenuate coefficient α. However, in both form conditions, omitted or non-reached items

were not scored incorrect but imputed along with planned missingness. Additionally, the amount of unplanned missingness was very small overall (see Table 5). Therefore, the disparity in reliability coefficients is not likely due to speededness.

Another explanation for the disparity in reliability estimates was explored. Attali (2005) noted that an increase in guessing behavior *on a few items* at the end of a test will mathematically deflate α. However, rapid guessing has been shown to be present throughout low-stakes testing contexts (S. L. Wise, 2006). Rapid guessing by some examinees on items throughout the test has been shown to artificially *inflate* coefficient α (S. L. Wise & DeMars, 2009). The reason is that rapid guessing behavior throughout a test can result in variance that appears systematic. This systematic variance increases inter-item correlations, and subsequently, coefficient α. If the Short Form condition resulted in a reduction in rapid guessing behavior, perhaps the Long Form data contains more rapid guessing and a spuriously inflated coefficient α. To assess this possibility, motivation filtering was used to remove cases with non-effortful responses or guesses (V. L. Wise, S. L. Wise, & Bhola, 2006). This same procedure was applied to the current study to explore this possibility. Table 10 contains the NW-9 average coefficient α values before and after filtering out cases with Test-taking Effort subscale scores < 15.

Table 10
NW-9 Coefficient α Before and After Motivation Filtering by Form Condition

	Before Motivation Filtering			After Motivation Filtering			
	N	Mean (SD)	α	Min *N*	Max *N*	Mean (SD)	α
Long Form	216	47.64 (8.65)	.860	175	178	49.24 (7.73)	.834
Short Form	1,027	49.02 (7.97)	.827	863	869	50.08 (7.31)	.801

Note. Motivation filtering removed participants who reported Test-taking Effort scores < 15. Sample sizes after motivation filtering vary due to the imputation of SOS subscale scores differing between imputed datasets. Test-taking Effort and SAT Math resulted in near zero correlations in both groups: $r = .01$ and $r = .11$ in the Long Form and Short Form, respectively.

If more rapid guessing was occurring in the Long Form condition than the Short Form as a result of low motivation, coefficient α would be expected to *decrease more* in the Long Form after filtering. Instead, motivation filtering simply reduced test score variance which reduced coefficient α values in both conditions. Therefore, the disparity in coefficient α estimates is likely due to less test score variance in the Short Form condition overall.

The simplest explanation for the small disparity in coefficient α values is the slight decrease in test score variance in the Short Form condition. One definition of reliability is the ratio of true-score variance to total variance ($\rho\rho' = \frac{\sigma_T^2}{\sigma_X^2}$). True-score variance ($\sigma_T^2$) is covariance between scale items whereas error variance is unique item variance; the sum of these quantities is total variance (σ_X^2). Coefficient α considers any covariance among items to be true-score variance, and when this is reduced, the ratio of true-score variance to total variance will decrease. For the purposes of the current study, introduction of the PMD did not unduly affect test reliability.

Item-Level

Unplanned missingness. Although 50% missingness in the Short Form was planned by design, some unplanned missing data was expected in both form conditions. Overall, the proportion of unplanned missingness was similar between the Long and Short Form conditions and very small. Over 99% of the sample responded to each item administered. Although small overall, the presence of unplanned missing data during the test was more pronounced in some areas. The percentage of students with omitted item

responses for each condition is depicted in Figure C3. These percentages were calculated to provide a fair comparison of unplanned missing data between the form conditions.

Surprisingly, unplanned missing data was more pronounced in the Short Forms. Readers may notice that missing data appeared to follow an interesting pattern with several "bumps" around items 17, 33, 50, and 66. Readers may recall that test forms were assembled by combining each pair of four approximately equal item sets (see Appendix A). Items 17, 33, 50 and 66 demarcate the end of the four item sets. The end of each item set is also the end of one or two Short Forms. Omitted items occurred more frequently at the end of some Short Forms suggesting that a few forms may be *slightly* speeded. Recall that 30 minutes was determined to be the testing time for all Short Forms, but students may have needed 35 minutes. A reduction in 50% of the test may not mean a reduction in 50% of testing time. Practitioners using the three-form design should time each form individually to determine if estimated testing time is sufficient to prevent speeded test forms and unnecessary missing data.

Proportion of students omitting items for the SOS was a bit more pronounced than the NW-9 in both form conditions (see Figure C3). Overall, SOS missingness was very small with less than 3% of the sample not responding to each item in both conditions. The missingness mechanism behind SOS omissions is unclear. One explanation could be that students completed a page of scientific reasoning items, flipped the page, and note items of a very different format. The top of the page reads: "Natural World-Student Opinion Scale" followed by new test instructions. It could be that students did not respond to these items as they perceived them to be unimportant or it could be that students simply missed this page as it was printed on the back of the last test page. The

notion of fatigue is not plausible here as the rate of SOS missingness was similar in both form conditions. With only about 43 items to complete in each Short Form condition, including the SOS, one would expect SOS missingness to be *lower* than in the Long Form condition if fatigue played any role. It remains unclear why this meager increase in missing data was observed. Perhaps the SOS measure should be administered immediately following the final test question so that students do not miss the scale.

Item parameters. Very few differences in item difficulty and discrimination parameters were observed between the Long and Short Form conditions. The small amount of DIF that was present did not seem to relate with item position so a fatigue effect was not inferred. Moreover, this significant DIF did not consistently favor one group over the other. More research should be conducted to compare these designs at the item-level. However, at this point, the study results suggest that assessment practitioners may not be concerned with difference in item-level quality when considering a PMD. The reader may wonder if the DIF comparison in this study between Long Form and Short was a fair comparison given the lack of a consistent order for each item in the Short Form condition. Recall that each item appeared in the Short Forms in an "early" order position (i.e., the first half of a Short Form) and in a "late" position. Some items appeared in an early or late position on more than one Short Form. For example, item 1 on the NW-9 appears at the very beginning of Short Forms A and E. However, with Short Form D, that same item appears as item 18. Early and late positions *within* the Short Form condition may result in position effects. Imputation after combining all Short Forms together may "muddle" existing differences. To investigate this possibility, item difficulties (p-values)

were calculated for each item by NW-9 Short Form *before imputation*. These values are displayed graphically for items 1-33 in Figure C4 and items 34-66 in Figure C5.

Long Form difficulties are denoted by a square symbol and the Short Form difficulties are denoted with a circle if that item was in the first half of the Short Form, or "early", and a triangle if that item was in the second half of the Short Form, or "late." If fatigue resulted in more difficult items, the same item would appear easiest in the early position and hardest in the Long Form, particularly for items appearing late. Short Form items with the late position presentation would most likely exhibit difficulty indices lying somewhere in between. Generally, item difficulties do not appear to be systematically different across forms. However, toward the end of the 66-item test (items 62-66), the differences in the difficulty indices appear a bit more pronounced. Note that items 62 through 66 appear easiest in the early position on the Short Form and hardest at the end of the Long Form. However, these differences are quite small and there is no consistent pattern of item difficulties for all items at the end of the test. If the test was even longer, or if the examinees were not as motivated, a greater effect may have been observed. This issue will be discussed further in a later section.

Examinee Motivation

The final two research questions examined differences in the examinee experience between the two form conditions. Specifically, mean self-reported Test-taking Effort and Test Importance variables were compared as well as the relationship between Test-taking Effort and test performance. Participants in the Short Form condition reported slightly higher Test-taking Effort (d = .115) and Test Importance (d = .116) than the Long Form condition. Although these mean levels were not statistically significant, I was pleased that

any observed difference was in the desired direction. Overall, students in both conditions reported relatively high motivation. Both subscale scores range from 5 to 25. The means for Test-taking Effort were around 18.5 to 20 or a 4 ("Agree" category) on the 5-point response scale for most items. Test Importance means were a bit lower, around 15, which is about a 3 ("Neutral" category) for each item. These relatively high examinee motivation group means may be a function of the assessment culture at the institution where this study was conducted. The presence of this culture is clear given the high attendance rate (≈ 90%), standardized proctor training, and test instructions that encourage effort. All of these procedures have shown increases in Test-taking Effort scores (Lau et al., 2009). Without these practices in place, the form conditions examined in this study may have resulted in a larger effect on motivation variables. This issue will be considered further in a later section of this discussion.

The relationship between self-reported Test-taking Effort scores as measured by the SOS (Sundre & Moore, 2002) and test performance on the NW-9 did not differ between the test form conditions. However, the relationship is in the desired direction for the Short Form condition. That is, the relationship between Test-taking Effort and test performance was smaller (r = .399) in the Short Form condition than the Long Form condition (r = .424). The correlations between Test-taking Effort and test performance in this study are similar in magnitude to those obtained in other low-stakes assessment contexts. Several researchers have used the SOS Effort subscale to measure test-taking effort and have calculated correlations that range from .251 to .450 across various tests and contexts. Other studies have used SOS Total scores (i.e., the sum of Test-taking Effort and Test Importance) to correlate with test performance. To provide a fair

comparison to the current study, studies that used Test-taking Effort alone were included in this review. See Table 11 for a sampling of these studies.

Table 11
Reported Correlations between SOS Effort Scores and Low-Stakes Test Performance

Study	Test	Pearson *r*
Current Study	Natural World-9 Long Form	.424
	Natural World-9 Short Form	.399
Abdelfattah (2010)	Math Test	.251
	Science Test	.255
Kornhauser, Minahan, Siedlecki, and Steedle (2014)	Collegiate Learning Assessment (CLA) Academic Citizenship Consequence Condition	.280
	Collegiate Learning Assessment (CLA) No Consequence Condition	.256
V. L. Wise, S. L. Wise, and Bhola (2006)	Information Literacy Test	.45
	Fine Arts	.33
	Natural World-6	.31
	American Experience	.35
	Sociocultural Dimension Assessment	.34
Waskiewicz (2011)	PCOA	.42

Note. All studies included in this table were conducted in low-stakes testing contexts, according to the authors. Some studies have used SOS Total scores (i.e., the sum of Test-taking Effort and Test Importance, such as Wolf and Smith, 1995) to correlate with test performance. However, these studies were excluded because subscale scores only were used in the current study.

Albeit non-significant, there was a slight decrease in the bivariate relationship between Test-taking Effort and test performance for the Short Form condition. This finding is in the desired direction when considering the deleterious effect of low motivation on test scores. That is, with less shared variance between Test-taking Effort and test performance, the Short Form condition is less influenced by a lack of effort by some examinees. In fact, the PMD resulted in a reduction in 2% of the variance shared between Test-taking Effort and test scores. This may be considered by some as a

reduction in construct-irrelevant variance in assessment test scores (Haladyna & Downing, 2004).

The question at this point becomes: how does an assessment practitioner know when test scores are free of construct-irrelevant variance from examinee motivation? Logic suggests that *some* effort is required for examinees to complete a test to the best of their ability. Therefore, we should expect a non-zero relationship between effort and test performance. However, at what point is this relationship concerning? And if we can decrease the magnitude, at what point will we be "satisfied" with this relationship? Perhaps the best comparison of this relationship is when the same test and effort measure are administered in a high-stakes context where a grade is tied to performance. What combination of interventions within a low-stakes context would result in similar effort and performance on the same test in a high-stakes context? It may be that we have already implemented the best practices we know and reducing examinee burden contributes little more to this quest for valid scores. The lack of larger effects on examinee motivation and other parameters may be due to the assessment culture at the institution where this study was conducted. The pervasive assessment culture limits the generalizability of this study to other institutions where assessment practice is not as well implemented and perhaps valued.

Study Limitations

As with any research design and study, limitations will exist and should be openly discussed and addressed with regard to study interpretations and generalizability. Limitations of the current study may have influenced the magnitude of the observed effects. These limitations are not always design flaws but rather commentary regarding

the scope of the study. For example, the test forms in the current study were administered first in the testing session. This test has traditionally been the first administered, and that placement was replicated to assure results comparability. This design was chosen to evaluate the possibility of fatigue within a single test rather than fatigue occurring after completing a battery of exams. However, a replication of this study which manipulates the length of the *last* test form in a series of tests could result in greater differences in both test and item parameter estimates. Studies that have examined data quality over a testing session would suggest that examinee burden would be greatest at the end of a testing session (DeMars, 2007).

The item-level differences observed between the forms may not be a large concern given that the six-form PMD was not fully balanced. That is, some items were given in about the same position on more than one form. This design may influence item-level parameters differentially. Even so, very few differences in item parameters were observed. These few differences may be due to context clues or priming differences within the forms. For example, one of the NW-9 Short Forms may have a series of related questions followed by some questions that are notably different and difficult. Item context was not considered when creating the Short Forms. Researchers who study PMDs, particularly the three-form design, would benefit from a more careful consideration of form balance in terms of item content and difficulty. Although the BIB emphasizes balance in item position and form length, several more forms are required to achieve this balance (see Table 1). With a fixed sample size, the three-form design will result in a larger *N* for each item covariance than the BIB, and this feature may influence the stability of parameter estimation using MI and ML techniques. However, large effects

in item parameter differences were not observed and what was observed did not favor one form condition over the other. Thus, balance may not be a large concern. The six-form design used in this study is still viable but with planned missingness set at 50%, it does introduce a large amount of missing data that could affect ease of estimation for practitioners.

Estimation and recovery of the Short Form condition missing data required changes to the default iteration criterion and the use of a ridge prior. These extra steps may have been avoided if another design with less missing data was chosen. The six-form design used in the current study was chosen to reduce testing burden and time by 50% rather than 33% in a three-form design without an X-set. With 66 item means and 2,145 item covariances to estimate, the EM algorithm took several iterations with 50% missing data. Estimation of the EM and data augmentation algorithms may have been faster with less missing data, but the desire to reduce examinee burden substantially resulted in the six-form design. That is, the six-form design was an attractive option to cut examinee burden in half. Although the six-form design is certainly a viable option, similar effects on test parameter estimates and motivation variables may be observed with a three-form design without increased burden on the practitioner. Moreover, less missing data may have eliminated the need to use a ridge prior.

A small ridge prior ($d^* = 10$) was used to assist imputation model convergence in the Short Form data. Ideally, a non-informative prior would have best suited this study to allow the data to represent the posterior distributions fully. This parameter was chosen according to suggestions by methodologists to reduce possible bias (Graham, 2012). To assess the influence of the ridge prior and enhance appropriate interpretation of results,

another Short Form imputation model was run with a d^* value of 5 instead of 10. All relevant parameter estimates to the original Short Form imputation model were nearly equal and the conclusions were substantively identical. For example, the correlation between Test-taking Effort scores and NW-9 scores was $r = .395$ when using a ridge prior of 5. This correlation is nearly identical to the one produced from the imputation model with a ridge prior of 10 ($r = .399$). This information coupled with the favorable degrees of freedom (v_1) and FMI values should reassure readers that the ridge prior simply helped the data augmentation algorithm without introducing sizeable bias.

The largest limitation to the generalizability of the results was the context within which the current study was conducted. As mentioned, the assessment culture at this institution was well-established with several practices in place to ensure accurate measurement of student learning. Even within this culture, notable differences were observed. Perhaps larger differences would have been observed on a sample of senior-level students at this institution rather than students mid-way through their educational career. Practitioners who find themselves at institutions without an established culture of assessment may expect substantively larger effects. It is within the context of institutions new to assessment that PMDs may prove most useful for future study of examinee motivation.

Suggestions for Future Study

The synthesis of planned missingness and examinee motivation literature is an area ripe for further study. Assessment practitioners are encouraged to explore the benefits of these designs in practice to meet their data collection needs. Researchers who are interested in this burgeoning area of assessment should consider these suggestions

when designing future studies. Areas of research discussed next encompass balance of PMD forms, appropriate pooling of reliability coefficients, and the use of testing software to ensure actual item content delivery and collect more information about the examinee experience.

As mentioned, the six-form design used as the PMD in the current study was not perfectly "balanced" in terms of item position or form difficulty or content. The item-level findings of the current study could serve as a springboard for future study in this area. Perhaps form balance is not a large concern but this area of research could contribute to the use of PMDs in educational assessment contexts.

Recall that parameter estimates are computed within each dataset separately and these estimates are pooled across imputations to incorporate uncertainty due to the imputation model. Methodologists in MI have suggested "normalized" transformations to parameter estimates prior to pooling in hopes of improving this calculation. For example, correlations are transformed to the z-score metric, averaged across imputations, then the final estimate is transformed back to the correlation metric for reporting. There is no suggested transformation for coefficient α at this time so the multiply-imputed estimates were not transformed prior to averaging. Simulation studies could explore whether this practice introduces bias to the pooling procedure or, if a transformation is preferred, what transformation is most appropriate.

PMDs work well for paper-and-pencil tests but in a world where assessments are increasingly administered electronically, testing software could ease administration of shortened test forms. Participants could receive a random subset of a fixed number of items or pre-planned forms that balance content and administration time. Electronic

delivery of test forms would help the researcher deduce the exact content of the items administered to each student. This would ease scoring procedures and ensure data quality rather than asking participants to indicate which form they are completing. Although proctors in the current study reiterated the importance of this Form ID item several times, they were not able to ensure that all students followed this instruction perfectly. We feel confident that our data was correctly gathered and coded; however, the ability to reliably conduct such research elsewhere may be quite discrepant. Future studies could explore the implementation of PMDs as computer-based tests to collect rich information about the examinee experience. Effort on each item could be assessed by response time, rather than self-report, and this information could further explicate the examinee's experience during shortened accountability tests (S. L. Wise & Kong, 2010). Response time on previous items may help explain unplanned missingness as either a function of speededness, fatigue, or low motivation. These suggestions for future study tie closely with implications for the use of PMDs in higher education as discussed in the next section.

Implications for PMDs in Higher Education

Assessment practice. This study supported the notion that test and item-level data quality was not jeopardized by the PMD implemented. These findings support those of many previous researchers. PMDs provide a reduction in examinee testing time while allowing for the accurate estimation of group ability parameters. Assessment practitioners who are limited on testing time or are concerned that examinees may find a long testing session to be burdensome should take solace in these findings. Through the use of these designs, practitioners may collect all assessment information that is desired or mandated

without requiring each examinee to complete all tests. Any PMD may be created by practitioners. If all item pairs are administered on at least one form and participants are randomly assigned to test form, any design could be used to best meet institutional assessment data collection needs. Practitioners should not have to choose between demands for accountability and assessment data and demands on examinee time. The balance between these two demands becomes attainable with PMDs.

A small effect in group mean test performance due to form design is very encouraging for practitioners assessing learning over time. That is, some higher education institutions assess the "value-added" to their students as a result of their college experience (Steedle, 2012). In fact, some state boards require each higher education institution to report value-added information (State Council of Higher Education in Virginia, 2007). Assessment practitioners responding to this recommended practice should then desire to accurately measure this change over time. Best practice in the assessment of change over time is a repeated-measures design on the same group of students (Castellano & Ho, 2013). The effect of form design observed in this study is additive when considering the assessment of learning over time. That is, students administered a PMD performed better than students administered a full-form which goes above and beyond learning gain attained from the college experience. Prior research has indicated that value-added estimates, as measured by change in test performance over time, are attenuated due to a negative change in examinee motivation (Finney et al., 2015). Given this predicament, assessment practitioners should be enthusiastic about any changes in practice that may counter this attenuation.

This study also illustrates the ease of modern missing data handling techniques to account for missing data. Assessment practitioners should not be concerned about losing statistical power when employing a PMD. Several statistical packages, many of which practitioners may use already, have MI or ML capabilities built-in that would allay statistical power concerns. As discussed, some of these resources are freely available with guides for implementation (Peng, Harwell, Liou, & Ehman, 2006). Although SAS may not be available to most practitioners, freeware such as NORM (Schafer, 1999) is available for download that can assist practitioners in using ML and MI techniques with stand-alone software or SPSS (http://methodology.psu.edu/pubs/books/missing). With modern missing data techniques readily available, most assessment practitioners should find that these resources benefit their practice. For example, practitioners should not feel constrained to use traditional missing data handling techniques, such as listwise deletion, that have known biases on parameter estimates and standard errors. When considering missing data in scored right-wrong assessments, missing responses do not need to be and probably should not be scored as incorrect. Several methodologists have shown that scoring omitted items as incorrect results in person and item parameter bias (DeMars, 2002; Finch, 2011a, 2011b; Lord, 1974). With advanced techniques readily available to handle this missing data, practitioners need not score omitted items as incorrect.

Coupled with these advanced techniques, PMDs provide a method to collect institutional data of similar quality to administering all items to all examinees. To ease estimation, practitioners have several choices. The three-form design incorporates less missing data than the six-form design used in the current study which would ease estimation using MI. Although the six-form design was deemed best to explicitly reduce

examinee burden, the three-form design may be less computationally intensive. If practitioners still desire to incorporate more missing data, prior information could be specified to assist the data augmentation algorithm. For example, if the same test is given annually and no cohort effects are suspected, prior year's data can serve as reasonable priors for PMD data. With careful use of prior information, practitioners should be assured that their inferences based on PMD data are accurate (see Graham, 2012).

Examinee motivation research. The PMD did not substantially improve examinee motivation during a routine institutional accountability testing session. Several implications for these findings are as follows. The notion of examinee motivation is still a pervasive and principle concern for practitioners. As discussed in the limitations, the culture of assessment may have limited the observed effect of examinee burden on effort; however, this component should remain in discussions of motivation theory (S. L. Wise & Smith, 2011). Examinee burden or cost may be not be a factor early in a testing session but may begin to affect examinees and data quality after hours of testing. Other studies have suggested that this may be the case (DeMars, 2007); therefore, further study should examine these effects during a situation with more extreme examinee burden. The aspect of examinee burden should receive more attention in future research.

Reducing examinee burden through PMDs adds to the discussion of interventions used to increase examinee motivation such as tying performance to grades and interpretive score feedback. However, PMDs are unique in that the participant is unaware of an intervention and instead simply completes the test form assigned. Given the limited viability of many of the motivational interventions discussed in Chapter 2, where does examinee burden fit in this conversation? Which intervention will move the needle of

examinee motivation independent of other interventions? Comparisons of these interventions should include examinee burden as a possible predictor of examinee motivation.

Again, the current study was conducted during a routine education assessment testing session at an institution with an established culture of assessment spanning nearly three decades. The high attendance rates during the session coupled with the relatively high self-reported test-taking effort and importance scores reflect the dedication of these students to task completion. Other institutions without a robust culture of assessment may find greater effects of PMDs on test performance and examinee motivation. Perhaps other institutions that observed greater issues with examinee motivation may serve as appropriate settings for future examinee burden research. Until then, the cost associated with completing a test should remain in the theoretical framework as there seems to be indirect evidence of its influence.

General Summary and Conclusion

PMDs have largely been discussed as alternative methods to improve data collection efficiency. In addition to improved efficiency, the current study demonstrated that test performance during a PMD was higher than with a full-form design. These findings support the possibility that PMDs provide higher quality data than a full-form design as some authors have stated (T. D. Little et al., 2014; Littvay, 2009). That is, a reduction in burden for the examinee may have reduced fatigue enough to slightly increase test performance. Moreover, there was no evidence of deleterious effects on parameter estimates typically used for assessment and accountability purposes. Group means and standard errors were not negatively affected by the PMD when using a

modern missing data technique such as MI. Data may be collected using a PMD on existing assessments to decrease testing time, reduce examinee burden, and provide psychometrically sound estimates of group ability.

As discussed, the use of these designs at other institutions may result in larger effects than those observed in the current study. Practitioners at institutions without an established culture of assessment could expect to see sizable differences in group mean test performance and motivation variables when implementing a PMD. With increased pressure to provide accurate estimates of group ability, practitioners should explore PMDs and use them more frequently in assessment practice and research. This area of research is perfectly primed for more exploration of these innovative designs that could provide much needed benefits to higher education stakeholders.

The current study bridged the gap between PMD and examinee motivation literature to propose examinee burden as an important facet in the discussion of higher education assessment. A six-form design was proposed that reduced examinee burden by 50% and resulted in a significantly higher estimate for group-level ability when compared to a full-form design. Small effects on internal consistency estimates and item-level information were also observed. Although practically small, the PMD form design increased self-reported motivation variables for randomly assigned samples. The results of this study advocate for the inclusion of examinee burden as a part of the conversation of examinee motivation in low-stakes contexts. Practitioners and researchers in higher education assessment should see increased use of PMDs in coming years. These innovative designs show promise as a vehicle to increase data collection efficiency,

address motivation concerns, and provide test score information that leads to more accurate inferences of student learning and institutional quality.

References

Abdelfattah, F. (2010). The relationship between motivation and achievement in low-stakes examinations. *Social Behavior and Personality: An International Journal*, *38*, 159-167.

Allison, P. D. (2002). *Missing data*. Thousand Oaks, CA: SAGE Publications, Inc. doi: 10.4135/9781412985079

Allison, P. D. (2005). Imputation of categorical variables with PROC MI. *Case Analysis*, *30*, 1-14.

Allison, P. D. (2009). Missing data. In R. E. Millsap & A. Maydeu-Olivares (Eds.), *The SAGE Handbook of Quantitative Methods in Psychology* (pp. 72-89). Thousand Oaks, CA: Sage Publications, Inc.

Arbuckle, J. L. (1996). Full information estimation in the presence of incomplete data. In G. A. Marcoulides & R. E. Schumacker (Eds.), *Advanced structural equation modeling* (pp. 243-277). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Attali, Y. (2005). Reliability of speeded number-right multiple-choice tests. *Applied Psychological Measurement*, *29*, 357-368.

Baraldi, A. N., & Enders, C. K. (2010). An introduction to modern missing data analyses. *Journal of School Psychology*, *48*, 5-37.

Barnard, J., & Rubin, D. B. (1999). Small-sample degrees of freedom with multiple imputation. *Biometrika*, *86*, 948-955.

Barron, K. E., & Hulleman, C. S. (2015). Expectancy-Value-Cost model of motivation. In *International Encyclopedia of the Social & Behavioral Sciences* (2nd ed.) Oxford: Elsevier Ltd.

Barry, C. L., Horst, S. J., Finney, S. J., Brown, A. R., & Kopp, J. P. (2010). Do examinees have similar test-taking effort? A high-stakes question for low-stakes testing. *International Journal of Testing*, *10*, 342-363.

Battle, A., & Wigfield, A. (2003). College women's value orientations toward family, career, and graduate school. *Journal of Vocational Behavior*, *62*, 56-75.

Baumert, J., & Demmrich, A. (2001). Test motivation in the assessment of student skills: The effects of incentives on motivation and performance. *European Journal of Psychology of Education*, *16*, 441-462.

Brown, J. M., & Gaxiola, C. (2010). Why would they try? Motivation and motivating in low-stakes information skills testing. *Journal of Information Literacy*, *4*, 22-36.

Buck, S. F. (1960). A method of estimation of missing values in multivariate data suitable for use with an electronic computer. *Journal of the Royal Statistical Society, Series B*, *22*, 302-306.

Cao, J., & Stokes, S. L. (2008). Bayesian IRT guessing models for partial guessing behaviors. *Psychometrika*, *73*, 209-230.

Castellano, K. E., & Ho, A. D. (2013). *A practitioner's guide to growth models*. Washington, DC. Retrieved from http://scholar.harvard.edu/files/andrewho/files/a_pracitioners_guide_to_growth_models.pdf

Childs, R. A., & Jaciw, A. P. (2003). Matrix sampling of items in large-scale assessments. *Practical Assessment, Research & Evaluation*, *8*(16). Retrieved November 30, 2013 from http://PAREonline.net/getvn.asp?v=8&n=16

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Earlbaum Associates.

Cole, J. S., & Osterlind, S. J. (2008). Investigating differences between low- and high-stakes test performance on a general education exam. *The Journal of General Education*, *57*, 119-130.

Cole, J. S., Bergin, D. A., & Whittaker, T. A. (2008). Predicting student achievement for low stakes tests with effort and task value. *Contemporary Educational Psychology*, *33*, 609-624.

College Board. (2014). *Trends in college pricing*. Retrieved from http://trends.collegeboard.org/sites/default/files/2014-trends-college-pricing-final-web.pdf

Collins, L. M., Schafer, J. L., & Kam, C. M. (2001). A comparison of inclusive and restrictive strategies in modern missing data procedures. *Psychological Methods*, *6*, 330-351.

De Ayala, R. J., Plake, B. S., & Impara, J. C. (2001). The impact of omitted responses on the accuracy of ability estimation in item response theory. *Journal of Educational Measurement*, *38*, 213-234.

Debeer, D., & Janssen, R. (2013). Modeling item-position effects within an IRT framework. *Journal of Educational Measurement*, *50*, 164-185.

Deci, E. L., & Ryan, R. M. (1985). *Intrinsic motivation and self-determination in human behavior*. New York: Plenum.

DeMars, C. E. (2002). Incomplete data and item parameter estimates under JMLE and MML estimation. *Applied Measurement in Education*, *15*, 15-31.

DeMars, C. E. (2007). Changes in rapid-guessing behavior over a series of assessments. *Educational Assessment, 12,* 23-45.

DeMars, C. E. (2009). Modification of the Mantel-Haenszel and logistic regression DIF procedures to incorporate the SIBTEST regression correction. *Journal of Educational and Behavioral Statistics, 34,* 149-170.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological), 39,* 1-38.

Eccles (Parsons), J., Adler, T., Futterman, R., Goff, S., Kaczala, C., Meece, J., & Midgley, C. (1983). Expectancies, values, and academic behaviors. In J. T. Spence (Ed.), *Achievement and achievement motivation* (pp. 75-146). San Francisco: Freeman.

Eccles, J. S., & Wigfield, A. (2002). Motivational beliefs, values, and goals. *Annual Review of Psychology, 53,* 109-132.

Eklöf, H. (2010). Skill and will: Test-taking motivation and assessment quality. *Assessment in Education: Principles, Policy & Practice, 17,* 345-356.

Enders, C. K. (2001). A primer on maximum likelihood algorithms available for use with missing data. *Structural Equation Modeling: A Multidisciplinary Journal, 8,* 128-141.

Enders, C. K. (2010). *Applied missing data analysis.* New York: Guilford Press.

Enders, C. K., & Bandalos, D. L. (2001). The relative performance of full information maximum likelihood estimation for missing data in structural equation models. *Structural Equation Modeling: A Multidisciplinary Journal, 8,* 430-457.

Feldt, L. S. (1969). A test of the hypothesis that Cronbach's alpha or Kuder-Richardson coefficient twenty is the same for two tests. *Psychometrika, 34,* 363-373.

Feldt, L. S., Woodruff, D. J., & Salih, F. A. (1987). Statistical inference for coefficient alpha. *Applied Psychological Measurement, 11,* 93-103.

Finch, W. H. (2011a). The impact of missing data on the detection of nonuniform differential item functioning. *Educational and Psychological Measurement, 71,* 663-683.

Finch, W. H. (2011b). The use of multiple imputation for missing data in uniform DIF analysis: Power and Type I error rates. *Applied Measurement in Education, 24,* 281-301.

Finney, S. J., Sundre, D. L., Swain, M. S., & Williams, L. M. (2015). The validity of value-added estimates from low-stakes testing contexts: The impact of change in test-taking motivation and test consequences. Manuscript submitted for publication.

Gottschall, A. C., West, S. G., & Enders, C. K. (2012). A comparison of item-level and scale-level multiple imputation for questionnaire batteries. *Multivariate Behavioral Research, 47,* 1-25.

Graham J. W. (2012). *Missing data: Analysis and design*. New York, NY: Springer.

Graham, J. W., & Schafer, J. L. (1999). On the performance of multiple imputation for multivariate data with small sample size. In R. Hoyle (Ed.), *Statistical strategies for small sample research* (pp. 1-29). Thousand Oaks, CA: Sage.

Graham, J. W., Cumsille, P. E., & Elek-Fisk, E. (2003). Methods for handling missing data. In J. A. Schinka & W. F. Velicer (Eds.), *Research methods in psychology* (pp. 87-114). New York: John Wiley & Sons.

Graham, J. W., Hofer, S. M., & MacKinnon, D. P. (1996). Maximizing the usefulness of data obtained with planned missing value patterns: An application of maximum likelihood procedures. *Multivariate Behavioral Research, 31,* 197-218.

Graham, J. W., Hofer, S. M., & Piccinin, A. M. (1994). Analysis with missing data in drug prevention research. *NIDA Research Monograph, 142,* 13-63.

Graham, J. W., Olchowski, A. E., & Gilreath, T. D. (2007). How many imputations are really needed? Some practical clarifications of multiple imputation theory. *Prevention Science, 8,* 206-213.

Graham, J. W., Taylor, B. J., & Cumsille, P. E. (2001). Planned missing-data designs in analysis of change. In L. M. Collins & A. G. Sayer (Eds.), *New methods for the analysis of change* (pp. 335-353). Washington, DC: American Psychological Association.

Graham, J. W., Taylor, B. J., Olchowski, A. E., & Cumsille, P. E. (2006). Planned missing data designs in psychological research. *Psychological Methods, 11,* 323-343.

Haladyna, T. M., & Downing, S. M. (2004). Construct-irrelevant variance in high-stakes testing. *Educational Measurement: Issues and Practice, 23,* 17-27.

Hardt, J., Herke, M., & Leonhart, R. (2012). Auxiliary variables in multiple imputation in regression with missing X: A warning against including too many in small sample research. *BMC Medical Research Methodology, 12,* 184.

Horton, N. J., Lipsitz, S. R., & Parzen, M. (2003). A potential for bias when rounding in multiple imputation. *The American Statistician, 57*, 229-232.

Huffman, L., Adamopoulos, A., Murdock, G., Cole, A., & McDermid, R. (2011). Strategies to motivate students for program assessment. *Educational Assessment*, *16*, 90-103.

Jakwerth, P. M., Stancavage, F. B., Reed, E. D., Champagne, A., Dabbs, P., & Hedges, L. (1999). *NAEP validity studies: An investigation of why students do not respond to questions*. Washington, DC: U.S. Department of Education, National Center for Education Statistics.

Johnson, E. G. (1992). The design of the National Assessment of Educational Progress. *Journal of Educational Measurement*, *29*, 95-110.

Kingston, N. M., & Dorans, N. J. (1984). Item location effects and their implications for IRT equating and adaptive testing. *Applied Psychological Measurement*, *8*, 147-154.

Kiplinger, V. L., & Linn, R. L. (1995). Raising the stakes of test administration: The impact on student performance on the National Assessment of Educational Progress. *Educational Assessment, 3*, 111-133.

Klein, S., Benjamin, R., Shavelson, R., & Bolus, R. (2007). The Collegiate Learning Assessment: Facts and fantasies. *Evaluation Review*, *31*, 415-439.

Knowles, E. (1988). Item context effects on personality scales: Measuring changes the measure. *Journal of Personality and Social Psychology, 55*, 312-320.

Kornhauser, Z., Minahan, J., Siedlecki, K., & Steedle, J. T. (2014, April). *A strategy for increasing student motivation on low-stakes assessments*. Paper presented at the

annual meeting of the American Educational Research Association (AERA), Philadelphia, PA.

Lau, A. R., Swerdzewski, P. J., Jones, A. T., Anderson, R. D., & Markle, R. E. (2009). Proctors matter: Strategies for increasing examinee effort on general education program assessments. *Journal of General Education, 58*, 196-217.

Lautenschlager, G. J., & Meade, A. W. (2008). AlphaTest: A Windows program for tests of hypotheses about coefficient alpha. *Applied Psychological Measurement, 32*, 502-503. doi: 10.1177/0146621607312307

Little, R. J. A. (1988). A test of missing completely at random for multivariate data with missing values. *Journal of the American Statistical Association, 83*, 1198-1202.

Little, R. J. A. (1992). Regression with missing X's: A review. *Journal of the American Statistical Association, 87*, 1227-1237.

Little, R. J. A., & Rubin, D. B. (1987). *Statistical analysis with missing data*. J. Wiley & Sons: New York.

Little, T. D., & Rhemtulla, M. (2013). Planned missing data designs for developmental researchers. *Child Development Perspectives, 7*, 199-204.

Little, T. D., Jorgensen, T. D., Lang, K. M., & Moore, E. W. G. (2014). On the joys of missing data. *Journal of Pediatric Psychology, 39*, 151-62.

Littvay, L. (2009). Questionnaire design considerations with planned missing data. *Review of Psychology, 16*, 103-113.

Liu, O. L., Bridgeman, B., & Adler, R. M. (2012). Measuring learning outcomes in higher education: Motivation matters. *Educational Researcher, 41*, 352-362.

Lord, F. M. (1974). Estimation of latent ability and item parameters when there are omitted responses. *Psychometrika, 39,* 247-264.

Meyers, J. L, Murphy, S., Goodman, J., & Turhan, A. (2012, April). *The impact of item position change on item parameters and common equating results under the 3PL model*. Paper presented at the Annual Meeting of the National Council on Measurement in Education (NCME), Vancouver, B.C.

Monahan, P. O., McHorney, C. A., Stump, T. E., & Perkins, A. J. (2007). Odds ratio, delta, ETS classification, and standardization measures of DIF magnitude for binary logistic regression. *Journal of Educational and Behavioral Statistics, 32,* 92-109.

O'Neil, H. F., Abedi, J., Miyoshi, J., & Mastergeorge, A. (2005). Monetary incentives for low-stakes tests. *Educational Assessment, 10,* 185-208.

O'Neil, H. F., Sugrue, B., & Baker, E. L. (1995). Effects of motivational interventions on the National Assessment of Educational Progress mathematics performance. *Educational Assessment, 3,* 135-157.

O'Neil, H. F., Sugrue, B., Abedi, J., Baker, E. L., & Golan, S. (1992). *NAEP TRP task 3a: Experimental motivation study*. Los Angeles, CA.

Peng, C.-Y. J., Harwell, M., Liou, S.-M., & Ehman, L. H. (2006). Advances in missing data methods and implications for educational research. In S. Sawilowsky (Ed.), *Real data analysis* (pp. 31-78). Greenwich, CT: Information Age.

Peugh, J. L., & Enders, C. K. (2004). Missing data in educational research: A review of reporting practices and suggestions for improvement. *Review of Educational Research, 74,* 525-556.

Popham, W. J. (1993). Circumventing the high costs of authentic assessment. *Phi Delta Kappan, 74,* 470-73.

Porter, S. R., & Umbach, P. D. (2006). Student survey response rates across institutions: Why do they vary? *Research in Higher Education*, *47,* 229-247.

Raghunathan, T. E., & Grizzle, J. E. (1995). A split questionnaire survey design. *Journal of the American Statistical Association*, *90,* 54-63.

Robitzsch, A., & Rupp, A. A. (2009). Impact of missing data on the detection of differential item functioning: The case of Mantel-Haenszel and logistic regression analysis. *Educational and Psychological Measurement*, *69,* 18-34.

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, *63,* 581-592.

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys.* New York: J. Wiley & Sons.

Rubin, L. S., & Mott, D. E. W. (1984, April). *The effect of the position of an item within a test on the item difficulty value*. Paper presented at the annual meeting of the American Educational Research Association (AERA), New Orleans, LA.

Schafer, J. L. (1997). *Analysis of incomplete multivariate data*. London: Chapman & Hall.

Schafer, J. L. (1999). *NORM users' guide* (Version 2). University Park: The Methodology Center, Penn State. Retrieved from http://methodology.psu.edu

Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods*, *7,* 147-177.

Schafer, J. L., & Olsen, M. K. (1998). Multiple imputation for multivariate missing-data problems: A data analyst's perspective. *Multivariate Behavioral Research, 33,* 545-571.

Schlomer, G. L., Bauman, S., & Card, N. A. (2010). Best practices for missing data management in counseling psychology. *Journal of Counseling Psychology*, *57*, 1-10.

Setzer, J. C., Wise, S. L., van den Heuvel, J. R., & Ling, G. (2013). An investigation of examinee test-taking effort on a large-scale assessment. *Applied Measurement in Education*, *26*, 34-49.

Shoemaker, D. M. (1973). A note on allocating items to subtests in multiple matrix sampling and approximating standard errors of estimate with the jackknife. *Journal of Educational Measurement, 10,* 211-219.

Silm, G., Must, O., & Täht, K. (2013). Test-taking effort as a predictor of performance in low-stakes tests. *TRAMES: A Journal of the Humanities & Social Sciences*, *17,* 433-448.

Smith, L. F., & Smith, J. K. (2002). Relation of test-specific motivation and anxiety to test performance. *Psychological Reports*, *91*, 1011-21.

Smith, L. F., & Smith, J. K. (2004). The influence of test consequence on national examinations. *North American Journal of Psychology*, *6*, 13-25.

Snyder, N. (2012). Data coaching: Measuring the effects of feedback on low-stakes test motivation. (Doctoral dissertation, Drexel University). Retrieved from: https://idea.library.drexel.edu/islandora/object/idea%3A3813

Socha, A., Swain, M. S., & Sundre, D. L. (2013, October). *Do examinees want their test scores? Investigating the relationship between feedback, motivation, and performance in low-stakes testing contexts.* Paper presented at the annual conference of the Northeastern Educational Research Association (NERA), Rocky Hill, CT.

State Council of Higher Education for Virginia (2007). *Guidelines for assessment of student learning*. Richmond, VA: SCHEV Task Force on Assessment. Retrieved from: http://www.schev.edu/Reportstats/2007AssessmentGuidelines.pdf

Steedle, J. T. (2012). Selecting value-added models for postsecondary institutional assessment. *Assessment & Evaluation in Higher Education*, *37*, 637-652.

Steinberg, L. (1994). Context and serial-order effects in personality measurement: Limits on the generality of measuring changes the measure. *Journal of Personality and Social Psychology*, *66*, 341-349.

Sundre, D. L. (1997, April). *Differential examinee motivation and validity: A dangerous combination*. Paper presented at the annual meeting of the American Educational Research Association. Chicago, IL.

Sundre, D. L. (1999, April). *Does examinee motivation moderate the relationship between test consequences and test performance?* Paper presented at the annual meeting of the American Educational Research Association (AERA), Montreal, Quebec, Canada.

Sundre, D. L., & Kitsantas, A. (2004). An exploration of the psychology of the examinee: Can examinee self-regulation and test-taking motivation predict consequential and

non-consequential test performance? *Contemporary Educational Psychology*, *29*, 6-26.

Sundre, D. L., & Moore, D. L. (2002). The Student Opinion Scale: A measure of examinee motivation. *Assessment Update, 14*, 8-9.

Sundre, D. L., Thelk, A. D., & Wigtil, C. (2008). *The Natural World Test, Version 9: A measure of quantitative and scientific reasoning, Test Manual.* Harrisonburg, VA.

Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement, 27,* 361-370.

Thelk, A. D., Sundre, D. L., Horst, S. J., & Finney, S. J. (2009). Motivation matters: Using the Student Opinion Scale to make valid inferences about student performance. *The Journal of General Education, 58,* 129-151.

Thoemmes, F., & Rose, N. (2014). A cautious note on auxiliary variables that can increase bias in missing data problems. *Multivariate Behavioral Research, 49,* 443-459.

Thorndike, R. (2007). Fisher's Z transformation. In N. Salkind (Ed.), *Encyclopedia of measurement and statistics*. (pp. 361-365). Thousand Oaks, CA: SAGE Publications, Inc. doi: http://dx.doi.org/10.4135/9781412952644.n175

van Buuren, S. (2012). *Flexible imputation of missing data*. Boca Raton: Chapman & Hall/CRC Press.

Wainer, H. (1993). Measurement problems. *Journal of Educational Measurement, 30,* 1-21.

Waskiewicz, R. A. (2011). Pharmacy students' test-taking motivation-effort on a low-stakes standardized test. *American Journal of Pharmaceutical Education, 75,* 1-8.

Wigfield, A., & Cambria, J. (2010). Students' achievement values, goal orientations, and interest: Definitions, development, and relations to achievement outcomes. *Developmental Review, 30*, 1-35.

Wigfield, A., & Eccles, J. S. (2000). Expectancy-value theory of achievement motivation. *Contemporary Educational Psychology, 25*, 68-81.

Wise, S. L. (2006). An investigation of the differential effort received by items on a low-stakes computer-based test. *Applied Measurement in Education, 19*, 95-114.

Wise, S. L. (2009). Strategies for managing the problem of unmotivated examinees in low-stakes testing programs. *Journal of General Education, 58*, 152-166.

Wise, S. L., & DeMars, C. E. (2005). Low examinee effort in low-stakes assessment: Problems and potential solutions. *Educational Assessment, 10*, 1-17.

Wise, S. L., & Kong, X. J. (2010). Response time effort: A new measure of examinee motivation in computer-based tests. *Applied Measurement in Education, 18*, 163-183.

Wise, S. L., & Smith, L. F. (2011). A model of examinee test-taking effort. In J. A. Bovaird, K. F. Geisinger, & C. W. Buckendal (Eds.), *High-stakes testing in education: Science and practice in K-12 settings* (pp. 139-153). Washington, DC: American Psychological Association.

Wise, S. L., Pastor, D. A., & Kong, X. J. (2009). Correlates of rapid-guessing behavior in low-stakes testing: Implications for test development and measurement practice. *Applied Measurement in Education, 22*, 185-205.

Wise, V. L. (2004). The effects of the promise of test feedback on examinee performance and motivation under low-stakes testing conditions. (Doctoral dissertation,

University of Nebraska-Lincoln). Retrieved from: http://digitalcommons.unl.edu/dissertations/AAI3131570/

Wise, V. L., Wise, S. L., & Bhola, D. S. (2006). The generalizability of motivation filtering in improving test score validity. *Educational Assessment*, *11*, 65-83.

Wolf, L. F., & Smith, J. K. (1995). The consequence of consequence: Motivation, anxiety, and test performance. *Applied Measurement in Education*, *8*, 227-242.

Wolf, L. F., Smith, J. K., & Birnbaum, M. E. (1995). Consequence of performance, test, motivation, and mentally taxing items. *Applied Measurement in Education*, *8*, 341-351.

Yamamoto, K. (1995). *Estimating the effects of test length and test time on parameter estimation using the HYBRID model*. *ETS Research Report Series*. Princeton, NJ. Retrieved from https://www.ets.org/research/policy_research_reports/publications/report/1995/hxqp

Yucel, R. M., and Zaslavsky, A. M. (2003, August). *Practical suggestions on rounding in multiple imputation*. Proceedings of the Joint American Statistical Association Meeting, Section on Survey Research Methods, Toronto, Canada.

Zis, S., Boeke, M., and Ewell, P. T. (2010). *State policies on the assessment of student learning outcomes: Results of a fifty-state inventory.* Boulder, CO: National Center for Higher Education Management Systems.

Zwick, R. (1991). Effects of item order and context on estimation of NAEP reading proficiency. *Educational Measurement: Issues and Practice*, *10*, 10-16.

Appendix A

Table A1
Test Form Content by Item Order

Item Order	Form A	Form B	Form C	Form D	Form E	Form F
1	FORM ID	FORM ID	FORM ID	FORM ID	FORM ID	FORM ID
2	1	18	34	51	1	51
3	2	19	35	52	2	52
4	3	20	36	53	3	53
5	4	21	37	54	4	54
6	5	22	38	55	5	55
7	6	23	39	56	6	56
8	7	24	40	57	7	57
9	8	25	41	58	8	58
10	9	26	42	59	9	59
11	10	27	43	60	10	60
12	11	28	44	61	11	61
13	12	29	45	62	12	62
14	13	30	46	63	13	63
15	14	31	47	64	14	64
16	15	32	48	65	15	65
17	16	33	49	66	16	66
18	17	34	50	1	17	18
19	18	35	51	2	34	19
20	19	36	52	3	35	20
21	20	37	53	4	36	21
22	21	38	54	5	37	22
23	22	39	55	6	38	23
24	23	40	56	7	39	24
25	24	41	57	8	40	25
26	25	42	58	9	41	26
27	26	43	59	10	42	27
28	27	44	60	11	43	28
29	28	45	61	12	44	29
30	29	46	62	13	45	30
31	30	47	63	14	46	31
32	31	48	64	15	47	32
33	32	49	65	16	48	33
34	33	50	66	17	49	--[b]
35	--[a]	--[a]	--[a]	--[a]	50	--[b]

Note. FORM ID asked students to indicate which form they were completing for scoring and item comparison purposes. Item numbers are the serial position of the items on the NW-9 Long Form.

[a] Forms A-D contained 34 NW-9 items total.

[b] Form F contained 33 NW-9 items total.

Appendix B

Natural World-Student Opinion Scale

Please think about the **test that you just completed**. Mark the answer that best represents how you feel about each of the statements below.

A = Strongly Disagree
B = Disagree
C = Neutral
D = Agree
E = Strongly Agree

1. Doing well on this test was important to me.
2. I engaged in good effort throughout this test.
3. I am not curious about how I did on this test relative to others.
4. I am not concerned about the score I receive on this test.
5. This was an important test to me.
6. I gave my best effort on this test.
7. While taking this test, I could have worked harder on it.
8. I would like to know how well I did on this test.
9. I did not give this test my full attention while completing it.
10. While taking this test, I was able to persist to completion of the task.

Note. Items 3, 4, 7, and 9 are reversed scored prior to scoring. The Test-taking Effort subscale score is the sum of items 2, 6, 7, 9, and 10. The Test Importance subscale score is the sum of items 1, 3, 4, 5, and 8.

Appendix C

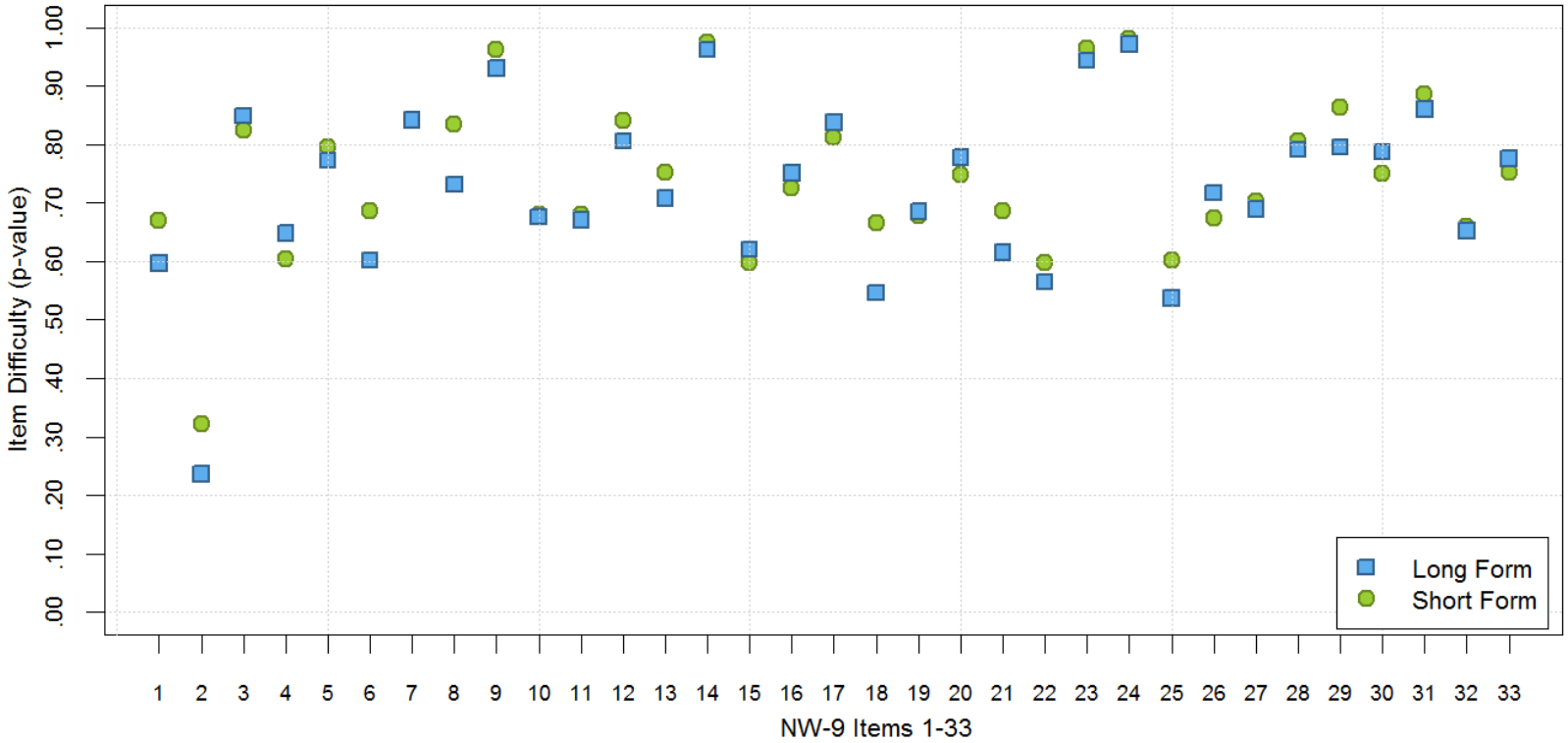

Figure C1. Item difficulty (p-values) for items 1 through 33 by form condition. Note that these difficulty parameters do not control for ability.

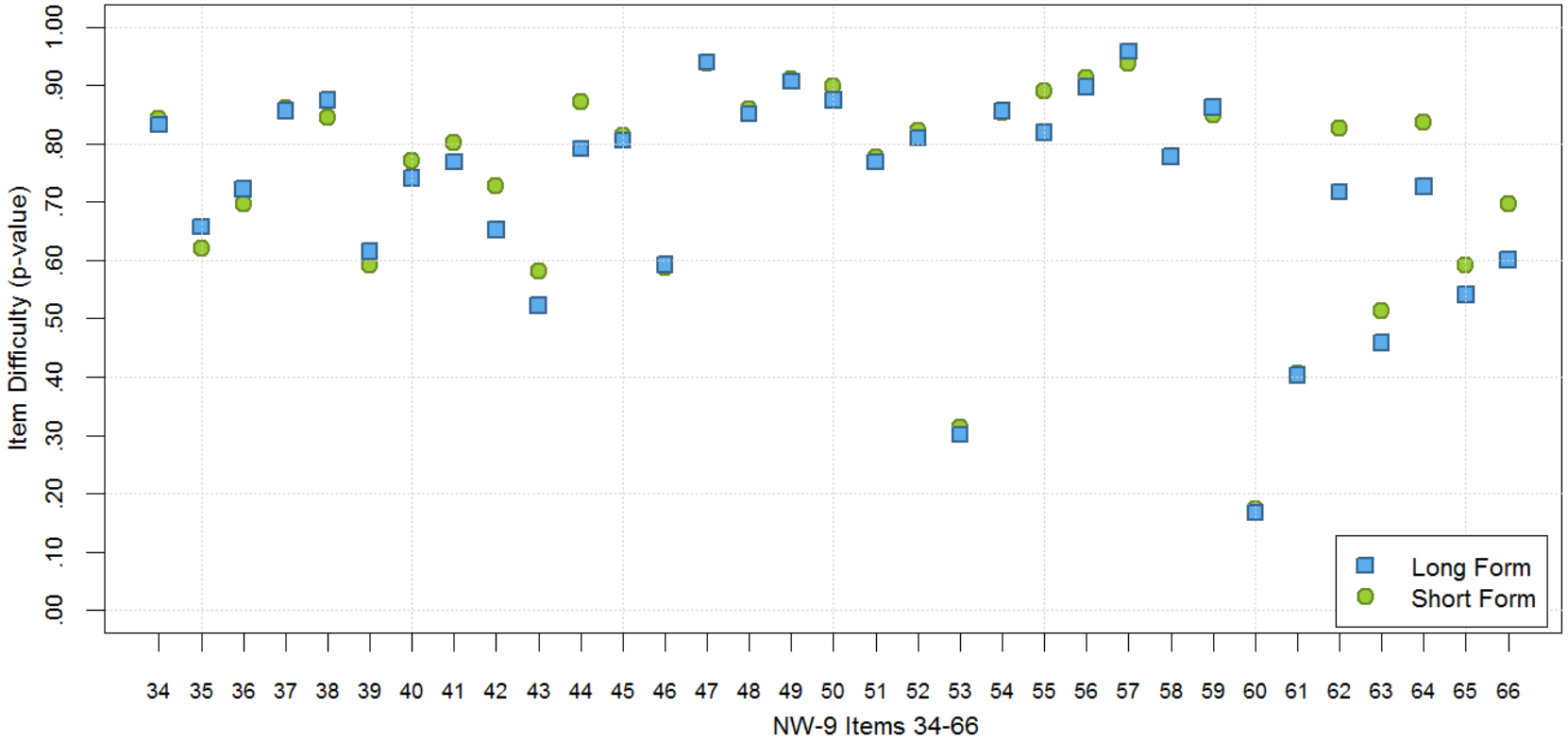

Figure C2. Item difficulty (p-values) for items 34 through 66 by form condition. Note that these difficulty parameters do not control for ability.

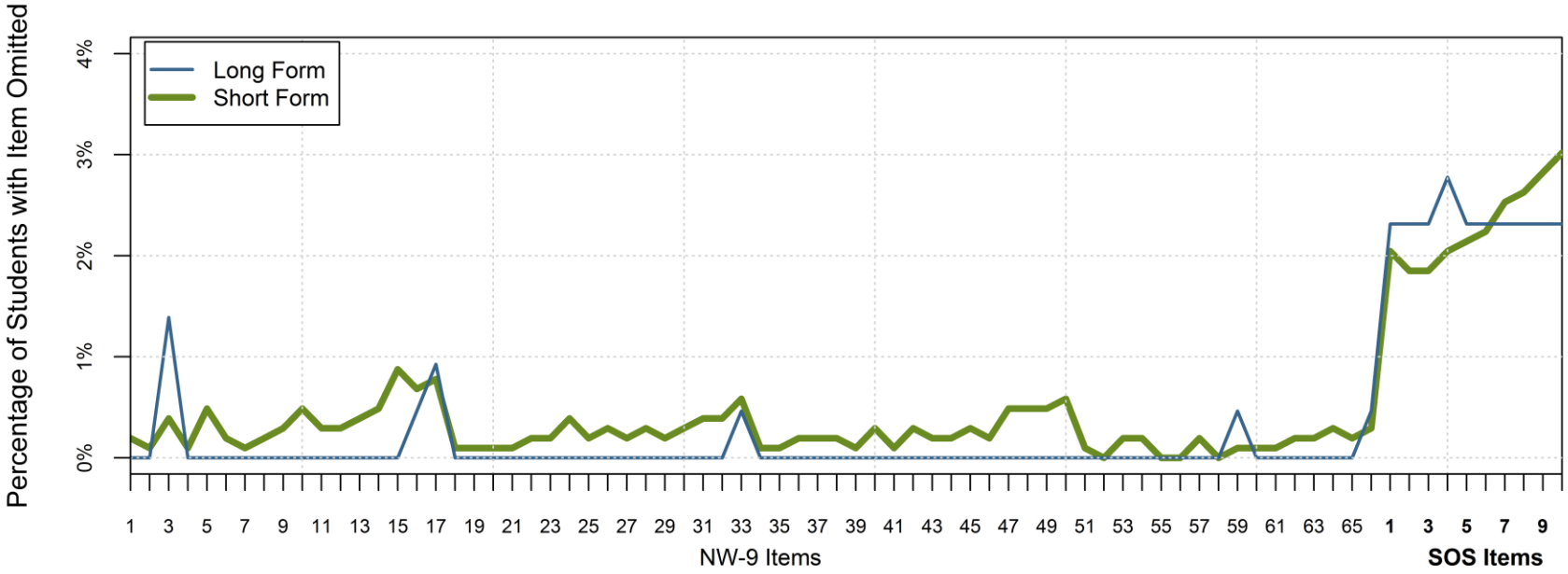

Figure C3. Percentage of student with item omitted. Short Form percentages do not include planned missing data. Note the slight increases in omitted items for the Short Form around items 17, 33, 50, and 66. These items were the final items in each item set (see Appendix A). SOS unplanned missingness overall was similar between the form conditions (see Table 5).

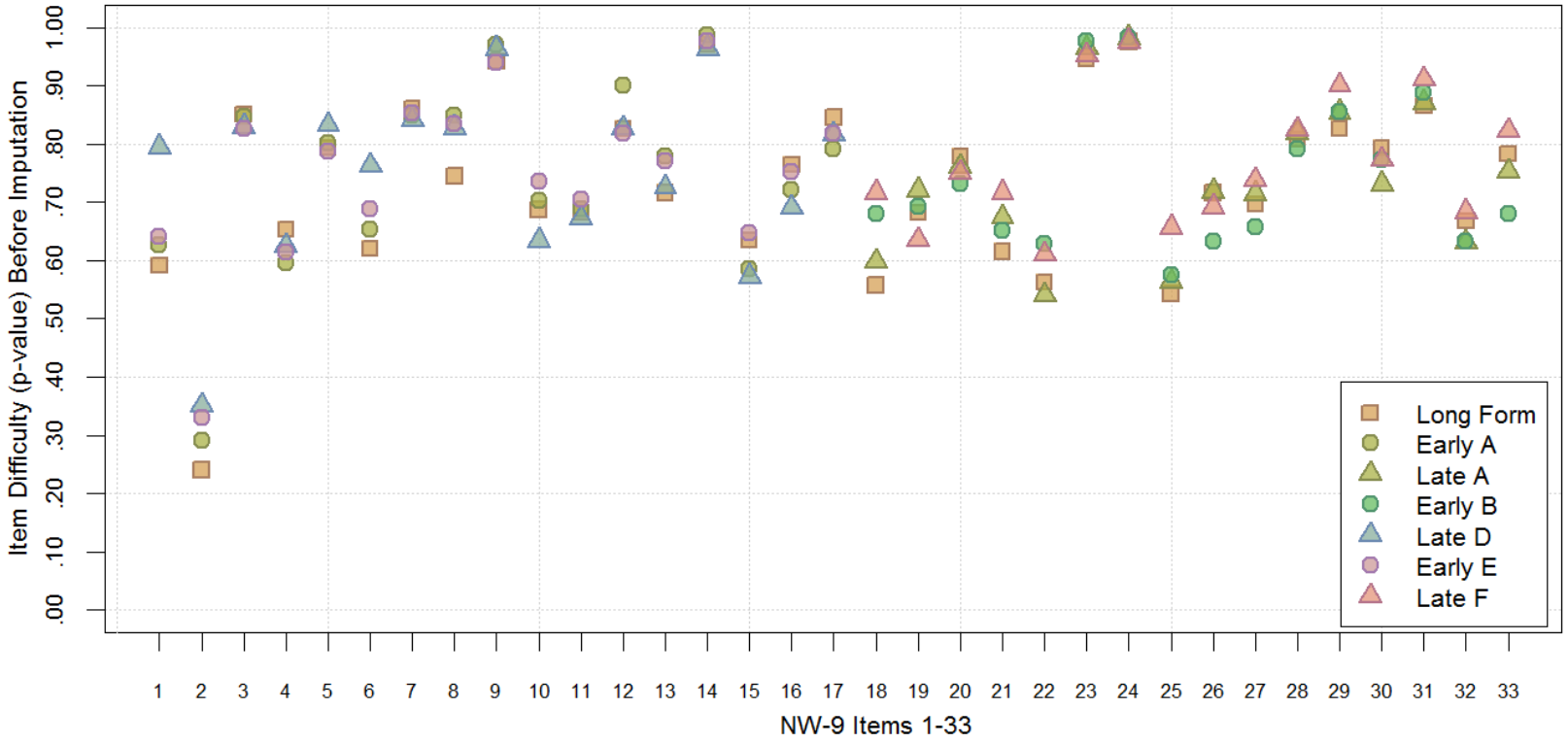

Figure C4. Item difficulty (p-values) for items 1 through 33 by form condition *before imputation*. Note that these difficulty parameters do not control for ability.

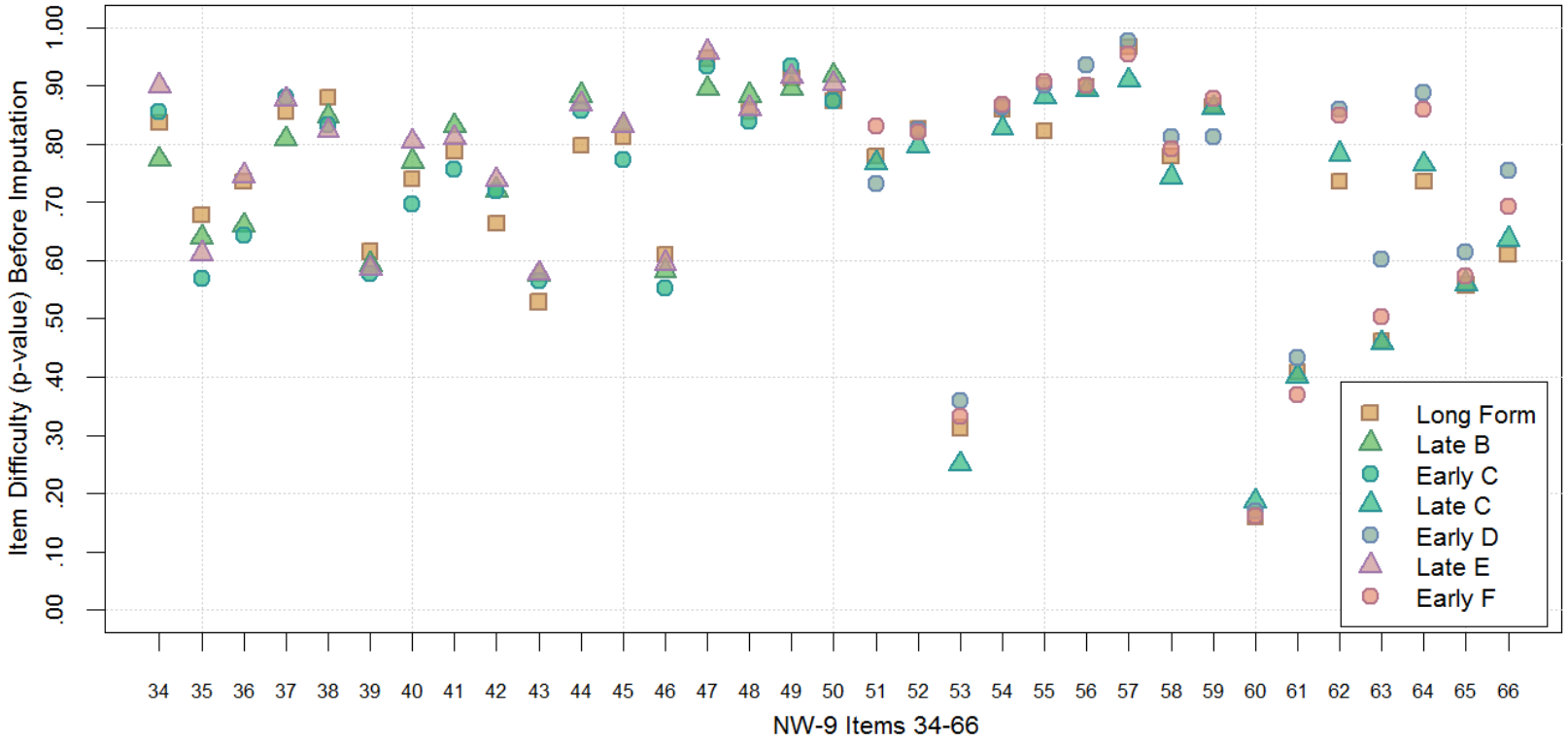

Figure C5. Item difficulty (p-values) for items 34 through 66 by form condition *before imputation*. Note that these difficulty parameters do not control for ability.